Анализирање на наклонетоста на медиумите преку надгледувано и ненадгледувано учење

Бојан Давитков

Факултет за информатички науки и компутерско инженерство Универзитет „Св. Кирил и Методиј“ Скопје, Република Северна Македонија

***Апстракт*-Во денешно време се повеќе се зголемува културната и политичката поларизација во општеството.Тоа се рефлектира на мас медиумите,а особено на порталите за вести,кои даваат раазлични,субјективни верзии на истите популарни вести.Луѓето поминуваат многу повеќе време за да се информираат за актуелните новости и моменталните случувања,па затоа е важно да бидат правилно известени.Секако дека постојат различни медиуми кои имаат свои агенди и точки на гледишта,од кои некои држат силен интегритет и се одличен приказ за новинарството,додека пак други не се изразуваат на таков начин.Јас сакав да пронајдам начин да стекнам увид во можноста за различни медиумски предрасуди за различни теми.Тоа го постигнав преку техники на машинско и длабоко учење, како и неколку интерактивни техники за визуелизација на зборовите кои се претходно претпроцесирани.**

Клучни зборови: Медиумска наклонетост - Длабоко учење - Машинско учење – Визуелизција- Bert- Корпус - Собирање на артикли

1.Вовед

Наклонетоста на медиумите ја претставува разликата во содржината или презентацијата на вестите[1]. Тоа е сеприсутен феномен во известувањето вести што може да има сериозно негативни ефекти врз поединците и општеството. Потенцијални проблеми за пристрасно покривање, без разлика дали преку изборот на теми или како тие се опфатени, е надополнето со фактот дека во многу земји неколку корпорации контролираат големи делови од медиумите. Дури и суптилни промени во зборовите што се користат во текстот од вестите може силно да влијаат на мислењата на читателите[2]. Во општествените науки, истражувањата во минатите децении резултирааа со сеопфатни модели за опишување на

пристрасност на медиумите, како и ефективни методи за анализа на медиумска пристрасност, како што е анализа на содржина[3].Но бидејќи тоа се изведувало рачно, не можело да го опфати огромниот број на вести кои се објавуваат во денешно време,а тоа не може да ни помогне да ги примениме изминатите истражувања на денешните типови на медиуми,кои се повеќе ги објавуваат вестите на своите веб-портали.

Откривање на медиумската пристрасност кон потрошувачите на вести, исто така, ќе помогне да се ублажат ефектите на пристрасноста и, на пример, да ги поддржат при носењето на поинформирани одлуки[4].Еден од предизвиците на овој проблем е да се создаде валиден класификатор на објавите на медиумите,а тој често зависи од достапноста на соодветниот корпус и квалитетот на собраните информации.Може да се каже дека важен дел од финалниот модел,особено кај делот од машинското учење кој се занимава со обработка на природните јазици, е изборот на атрибути,бидејќи раазлични техники од оваа област даваат различни резултати.Несоодветен избор на атрибути не само што нема да помогне во класификацијата, може да наштети и да ја намали точноста на класификаторот.

Во понатамошниот дел од ова истражување најпрво ќе се осврнам на сродните истражувања кои ме мотивираа да се посветам на овој проект,односно да се потрудам што е можно подетално да навлезам во анализата на наклонетоста во медиумите преку сеопфатно множество на техники кои ги обработуваат податоците од сите можни аспекти.Потоа, ќе се задржам на опис на системот,техниките и методите кои беа употребени за да се постигнат резултатите од ова истражување.Понатаму,во следната секција ќе бидат прикажани добиените резулататите споредени меѓусебе и со другите сродни истражувања на слични теми.Финално,на крајот на овој документ,ќе дадам краток осврт на овој проект заедно со насоките за идни истражувња и можностите за понатамошно подобрување на перформансите на користените технологии.

2.Сродни истражувања

Постојат голем број на студии на оваа тема,а бидејќи таа станува се поактуелна со развојот на технологијата,како и со влијанието на КОВИД-19 на известувањата на медиумите,тие се повеќе ги користат најновите пристапи во областа на машинското и длабокото учење за различни анализи на тема наклонетост и пристрастност на медиумите при нивното пренесување на вестите.

Едно од тие истражувања е анализата со помош на Латентна алокација на Диришле(LDA) за анализа на онлајн медиумите[5],каде авторите предложуваат комбиниран,унифициран модел со помош на повеќе техники од областа на NLP.Тие користат ЛДА за да го извлечат множеството теми од нивната база на податоци што опфаќа одреден месец за да имаат временска рамка за споредба и ги анализираат написите за вести кои потекнуваат од

индивидуални извори, а потоа прават општа екстракција на темата за ЛДА користејќи ги сите податоци во сите извори за тој месец. Потоа генерираат 10 теми за секој извор на вести и секоја ја карактеризираат со потпис на темата што ги содржи првите 10 зборови за темата. Темите на кои тие добиле најприфатливи резултати се однесувале на Кина, WikiLeaks и спорот околу двете Кореи.

Во следната студија,авторите[6] ја тестираат објективноста и непристрастноста на медиумите во Јужна Африка.Тие тврдат дека од уредниците се бара да одржуваат објективен и балансиран став без никаква корист за надворешните политички или корпоративни интереси. Оваа претпоставка за објективност ја тестираат во голем обем со пресметковна анализа на 30 000 написи објавени од пет медиумски куќи: News24, SABC, EWN, ENCA и IOL. Користејќи моделирање на теми, 38 теми се извлечени од корпусот и се пресметува сентимент за секоја тема. Студијата нагласува различни случаи и на прекумерно и недоволно известување од страна на медиумските куќи за одредени теми. Исто така, тие идентификувале разни пристрасности на тоналитет по медиумски куќи.

Ајвер et. al. [7] користат рекурзивни невронски мрежи за да ја откријат политичката идеологија на медиумите и нивните натписи. Земајќи ја предвид хиерархиската природа на природниот јазик, RNN можат да моделираат семантички состав,што е принцип дека значењето на фразата е комбинација на значењето на зборовите внатре во таа фраза и синтаксата што ги комбинира тие зборови. Иако семантичкиот состав не се применува универзално (на пример, сарказам и идиоми), повеќето од природните јазици го следат овој принцип.Тие откриле дека повеќето идеолошки пристрасности стануваат идентификувани само на повисоко ниво од нивоата на дрвја со реченици, па затоа моделите што се потпираат првенствено на статистика за распределба на ниво на зборови не се пожелни за тој тип на проблем.

Ахмед и Ксинг[8] користат ЛДА за анализа на темите од идеолошка перспектива на нештата. Нивниот модел, mviewLDA, го гледа секој документ како резултат на интеракцијата помеѓу неговите актуелни и идеалошки димензии. Моделот се обидува да ги објасни лексичките варијабли во документот со припишување на овие варијаблина една од тие димензии или на нивните интеракции. Тематските модели, како ЛДА, дефинираат генеративен процес за колекција на документи базирани на множество параметри. Тие користат три основни модели. Првиот основен модел е класификатор SVM обучен за нормализирана фреквенција на зборови на секој документ. SVM е трениран користејќи параметар за регулација во опсегот {1, 10, 20, · · ·, 100} и пријавен е најдобриот резултат (без извршена вкрстена валидација). Другите два се надгледувани модели на ЛДА: надгледувани ЛДА(sLDA) и discLDA,кој е условен модел што го дели достапниот број теми во теми специфични за класа и заеднички теми.

3.Опис на системот

3.1. Податочно множество и претпроцесирање

Податочното множество кое се користеше за овој проект се состои од прибрани записи од онлајн портали на неколку медиумски куќи.Тоа го направив со помош на newspaper3k библиотеката која ја нуди Python[9].За потребите на проектот ги искористив сајтовите на 12 медиумски куќи и со помош на гореспоменатата библиотека лесно ги презедов.На секој од изворите на податоци,односно порталите,е поставен лимит од 300 артикли по извор за да не се преземаат премногу артикли од еден извор,а помалку од другите,и за да не преземаме нерелевантни и непотребни артикли од некои извори.Што се однесува до преземањето,од секој артикл го земаме неговиот наслов,текстот и URL од неговиот вебсајт, односно изворот.Овие податоци се сместуваат во Python dataframe за понатмошна употреба. Како резултат на ова,имаме собрано 2075 артикли во нашето податочно множество.

За другиот дел од проектот,ги преземав рејтинзите за наклонетоста на медиумите од AllSides сајтот,кои обезбедуваат одделни оценки за пристрасност за вести и содржини за уредници/мислења за голем број медиуми. Сите оценки за пристрасност се базираат на онлајн, пишана содржина, а не емитувана, ТВ или радио содржина. Нивите оценки се флуидни и подложни на промени со текот на времето, бидејќи се собираат нови информации и се менуваат предрасудите. AllSides користи повеќепартиска, научна анализа за да ја оцени пристрасноста. Нивната методологија е научна, но секој поединец ќе има субјективно мислење за пристрасноста на секој даден извор[10].Во посебна dataframe табела на секој од артиклите според изворот му е доделен рејтинг кој го преземен од табелата на AllSides.Во нашето множество вредностите на рејтинзите се “Center”,што значи дека медиумот има релативна непристрастност,”Lean Left”,дека е наклонет кон либерални ставови,додека пак “Lean Right” означува дека е наклонет кон конзервативни ставови. ”Right”,пак означува дека медиумот има строго конзервативен поглед на нештата.

Како мерки на претпроцесирање на множеството,се изведени бришење на дупликатите на артиклите кои имаат идентичен нааслов или текст за да немаме вишок на слични или потполно исти вести.Исто така,ги чистиме вредностите кои се null или Nan,соодветно за да немаме нецелосни податоци во нашето множество.Друг метод на претпроцесирање кој е извршен е агрегација на изворите на податоци.Односно,ако сме земале артикли од повеќе стрници на еден сајт,како што се “Health” или пак “Politics” делот од,на пример, “CNN” наместо како извор да имаме податоци како што се “cnn.com/politics” или “cnn.com/health” ние то го агрегираме во основниот сајт за да може ефективно да ја примениме вредноста за наклонетост која ќе ја прочитеме од AllSides сајтот.

Потоа,на график ги имаме прикажано бројот на артикли кои сме ги презеле и претпроцесирале од секој извор за да водиме сметка каква ни е распротранетоста и пропорционалноста на истите.Како надополнување,во новата податочна структура се наоѓа и колона која ни се однесува на text_length,односно бројот на зборови за секој од текстовите

на артиклите.На крајот од овој дел,направено е и групирање според вредноста на медиумската пристрасност за да имаме преглед на тоа колку артикли имаме со соодветната медиумска наклонетост како вредност.

3.2. Latent Dirichlet Allocation(LDA)

Како еден од главните делови од проектот имав храбра идеја да изведам метод за машинско учење без надзор за да најдам кластери со различни теми во мојот корпус, на кои потоа ќе направам анализа на чувства и анализа на субјективност за да видам дали тоа ќе ми обезбеди интересни сознанија. Во суштина, требаше да направам модел Topic modelling за LDA, што е алгоритам за кластеризација без надзор, специјално користен во обработката на природниот јазик.Всушност,Topic modelling идентификува теми од анализа на збирка документи, забележува обрасци на зборови во нив и автоматски ги групира овие обрасци што најдобро опишуваат збир на документи.

LDA е техника за машинско учење која еволуираше од претходниот модел наречен Веројатна латентна семантичка анализа [6] (pLSA) за намалување на димензионалноста на одреден текстуален корпус, истовремено зачувувајќи ги неговите интензивни статистички карактеристики. ЛДА претпоставува дека секој документ во корпусот може да се опише како мешавина од повеќе латентни теми, кои, пак, се дистрибуции за зборовите пронајдени во документите на корпусот. ЛДА претпоставува дека документите се составени од листи на зборови каде редоследот на зборовите не е важен, односно bag of words пристапот . За да се генерира корпус од D документи, каде што секој документ има Nd зборови, и за вкупно T теми, генеративниот алгоритам на LDA е:

1. Избери N ∼ Poisson(ξ).(Големината на секвенца од зборови)

2. Избери θ ∼ Dir(α).

3. За секој од N-те зборови w_n:

(a) Избери тема z_n ∼ Multinomial(θ).

(b) Избери збор w_n од $p(w_n|z_n,\beta)$, мултиномијална веројатност условно зависна од z_n.

Мора да се напомене дека димензијалноста **k** на Диришлевата дистрибуција,а со тоа и променливата на тема z мора однапред да бидат познати и фиксни.Исто така,треба да ги параметризираме зборовните веројатности β за кои сметаме дека се фиксна количина која треба да се процени со алгоритмот.

За ова истражување се искористени 2 верзии на модели на LDA-по 1 од GenSim и Sklearn библиотеките соодветно.Пред да бидат пуштени на LDA моделите,текстовите се стемирани со исчитени од стоп зборови,стемирани со помош на PorterStemmer и токенизирани на зборови со помош на токенизерот кој ни го нуди spacy.

После токенизирањето на нашиот влезен текст,клучно е да најдеме биграми и триграми во нашиот корпус. Биграмите се два збора што често се среќаваат заедно во документот, а

триграмите се 3 збора што често се среќаваат.**Phrases** моделот на Gensim[11] може да изгради и имплементира биграми, триграми, квадграми и многу повеќе. Двата важни аргументи за Phrases моделот се *min_count* и *threshold.*Следно го креиме нашиот bag of words корпус со dictionary функцијата од *genism.corpora.* Gensim создава уникатен id за секој збор во документот. Произведениот корпус пресликување на (word_id, word_frequency). Во Gensim моделот по неколку експерименти се одлучив да ги искористам default вредностите з сите хиперпараметри и да го пробам моделот на 20 теми.

За визуелизација на Topic Modelling,ја користев библиотеката pyLDAvis,која овозможува интерактивна визуелизација на темите[12].Интересното кај овој начин е тоа што со едноствен интерфејс може да ги прегледме сите теми,заедно со зборовите кои најмногу се појавуваат во нив.Приказ на оваа визуелизација е прикажан во секцијата за резултати. Кај SKlearn моделот,исто така е искористена LDA со 20 теми.

Дополнително,кај делот од податочното множество каде на артиклите им беа доделени нивните вредности за медиумска наклонетост според изворот од кој се преземени. Исто така,во овој случај го искористив моделот на LDA обезбеден од страна на SkLearn. Овојпат направив 3 визуелизации.Податочно множество го поделив на 3 дела,така што:

1.Во првиот дел се наоѓаат колоните кои имаат Bias вредност “Center”.

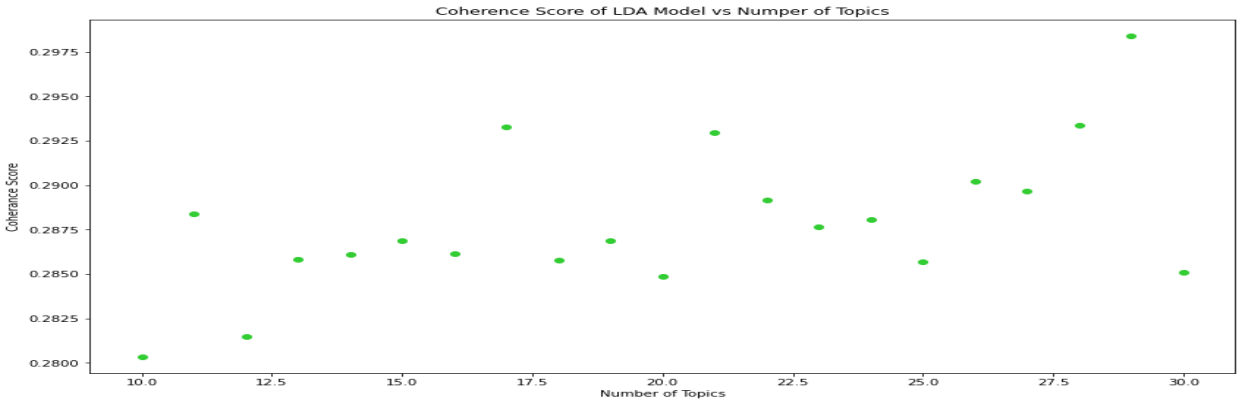
2.Во вториот дел се наоѓаат колоните кои имаат Bias вредност “Lean Left”.

3.Во третиот дел се наоѓаат колоните кои имаат Bias вредност “Right” или “Lean Right”.

За секој од овие делови направив LDA модел со 20 теми и визуелизација со помош на pyLDA библиотеката.

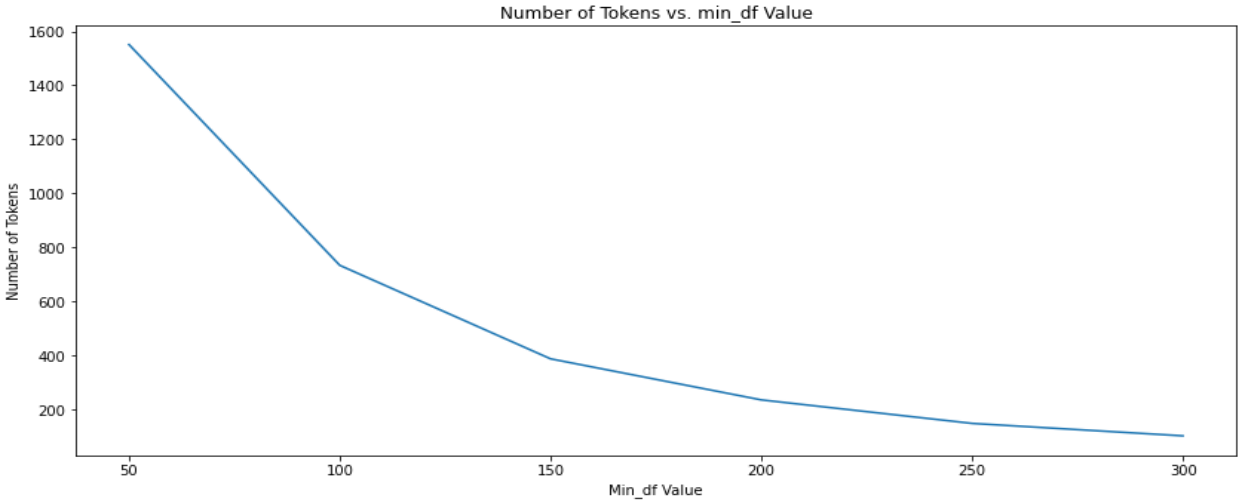
3.2.1.Прилагодувања

Како да знаеме дали нашиот модел со 20 теми извршил добро групирање на зборовите во нашиот корпус? За да донесеме одлука,треба да се најде објективна мерка за квалитетот. Како мерка за евалуација и прилагодување на моделот користев Кохерентност на тема (Topic coherence).Topic coherence го мери резултатот на една тема со мерење на степенот на семантичка сличност помеѓу зборовите со високи оценки во темата. Овие мерења помагаат да се направи разлика помеѓу теми кои се семантички толкувани и теми што се артефакти на статистички заклучок[13]. Со извлекување на кохерентноста на темата, би можеле да направиме хиперперметарско подесување за да го најдеме оптималниот број теми што треба да ги бараме! Бидејќи учиме без надзор, го пуштаме моделот во темнина со 20 теми и чекаме што ќе излезе.Во склоп на темата,го извлеков овој скор за моделите од 10 до 30 теми за да видиме кој ќе даде најдобро резултати,односно највисока оценка за кохерентност.Графикот од ова прилагодување може да го видиме на сликата.



Слика 1. Приказ на графикот за кохерентност на темата.

Кај LDA SkLearn моделот,се одлучив да направам експеримент со подесување на вредноста на *min_df* параметарот кај TF-IDF векторизерот.Што претставува овој параметар?При градење на вокабуларот се игнорираат терминте што имаат фреквенција на документи строго пониски од дадениот праг. Оваа вредност се нарекува и cut-off во литературата[14]. Ако се наоѓа во опсег од [0,0, 1,0], параметарот претставува дел од документите, целосни апсолутни броеви. Во случајот експериментирав со вредности на min_df од 50 до 300 (децимално од 0.05 до 0.3) на интервал од 50.Потоа на график го претставив бројот на токени наспрема вредноста на парметарот за да ја одредам вредноста каде што има најголемо израмнување.Ова е прикажано на следниот график.



Слика 2. Приказ на графикот за вредноста на min_df

3.3. Визуелизација

Следно,со помош на Scattertext библиотеката во Python[15] се одлучив да направам интересна визуелизација на податочното множество.Ова треба да ги илустрира разликите помеѓу левичарските и десничарските медиуми. Визуелизацијата прикажува клучни теми што се од интерес за двете страни, како и теми што се користат почесто од едната или

другата страна.За да се направи оваа визуелизација за разликите меѓу зборовите во корпусот,најпрво треба да се изврши стандардно чистење на текстот со отстранување на стоп зборовите и интерпункциските знаци.Потоа следува парсирањето на документите во зборови и со помош на функцијата од Scattertext модулот CorpusFromParsedDocuments го добиваме нашиот корпус од зборови.

Но,како да откриеме колку зборовите се асоцирани кон категориите?Scattertext ни дава можност да ја искористиме техниката Скалиран F-Score.Интуитивно, поврзаните термини имаат релативно висока прецизност специфична за категоријата и фреквенција на термини специфични за категоријата (т.е., % од терминте во категоријата се термини). За даден збор w_i во W и категорија c_j во C, ја дефинирме прецизноста на зборот w_i wrt во категорија како:

$$(w_i, c_j) = \frac{\#(w_i, c_j)}{\sum_{c \in C} \#(w_i, c)}.$$

Формула 1. Пресметување на прецизноста на збор во категорија

Функцијата #(w_i, c_j) претставува или колку пати w_i се јавува во документ означен со категоријата c_j или број на документи означени како c_j кои содржат w_i. Слично, дефинирајте ја фреквенцијата на збор што се појавува во категоријата како:

$$(w_i, c_j) = \frac{\#(w_i, c_j)}{\sum_{w \in W} \#(w, c_j)}.$$

Формула 2. Пресметување на фрекфенција на збор во категорија

Скалиран F-Score од овие 2 вредности е дефиниран како:

$$F_\beta(\ ,\ ) = (1+\beta^2)\frac{\cdot}{\beta^2 \cdot +}.$$

Формула 3. Пресметување на F-score

,каде што β припаѓа на R^+ и е фактор на скалирање каде фреквенцијата е фаворизирана ако β<1,прецизноста ако β>1, а се еднакво вреднувани ако β=1

Во случајот,ова го направив како пример за вредност на Bias==”Lean Left” и ги открив зборовите кои најмногу се појавуваат во оваа категорија.Анализа на ова ќе погледнеме во секцијата Резултати.

Финално,со помош на produce_scattertext_explorer функцијата,можеме на график да ги видиме зборовите кои се појавуваат во соодветниот корпус за дадена категорија.За проектот категориите беа “democratic” наспроти “republican”, бидејќи тоа е тема која често беше обработувана во медиумите од кои ги преземав податоцитеи сакав да ги воочам разликите во известувњето на медиумите за оваа тема.Графиците се направени за 3 вредности на медиумска пристрасност во нашето податочно множество(освен “Center”).

3.4. Класификација

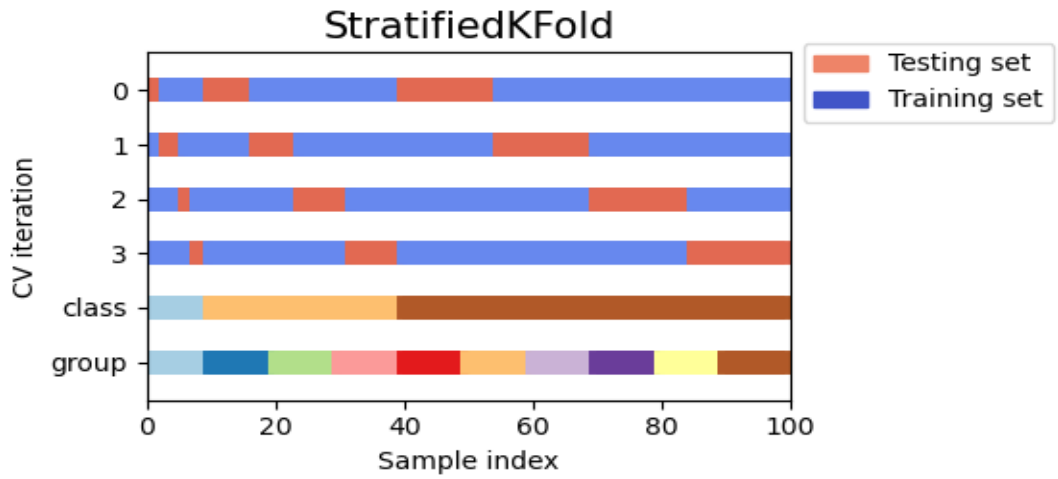
Како дел од овој проект,се обидов да направам 2 модели на класификација на податочното множество кое ги содржи и рејтинзите за медиумска наклонетост. Како мерки на претпроцесирање,ги тргнав сите артикли кои содржат помалку од 20 и повеќе од 3000 зборови бидејќи нивниот број беше значително мал,а може да се случи нивната инклузија да не ни даде добри резултати.

За двата модели го искористив LinearSVC класификаторот(Linear Support Vector Classification) од SkLearn.svm библиотеката. Слично на SVC со параметар кернел = 'линеарен', но имплементиран во линеарна смисла, а не смисла на Support Vector Machines, така што има поголема флексибилност во изборот на казни и загуби и треба подобро да се скалира на голем број примероци.Оваа класа поддржува густ и редок влез, а поддршката за повеќе класи се ракува според шемата „еден наспроти сите"[18].

Што се однесува до првиот модел,како атрибути беа искористени текстот и изворот,а излезот беа четирите Bias класи.Моделот беше збогатен со користење на Pipeline, така што во делот за текстот се инкорпорираше TF-IDF векторизер со биграми и триграми,а во делот за изворот се инкорпорираа биграми,триграми,4-грами и 5-грами.За тежините во трансформаторот, кои се мултипликативни тежини за карактеристики по трансформатор, по серија прилагодувања,одлучив да ги подесам на 0.5 за двата параметри.Множеството за тестирање беше подесено на 20% од вкупниот датасет.

Кај вториот модел пак,одлучив да го изоставам изворот и како атрибути да ги користам текстот,бројот на реченици и ARI индексот.И овој модел беше збогатен со користење на Pipeline, така што во делот за текстот се инкорпорираше TF-IDF векторизер со биграми и триграми.Овојпат тежините на трансформаторот бе подесени на 0.8 за атрибутот Text, 0.4 за атрибутот Sentences и 1 за атрибутот Reading.Повторно,множеството за тестирање беше подесено на 20% од вкупниот датасет.

Дополнително,на вториот модел по тренирањето и тестирањето се одлучив да извршам методи на крос-валидација,поточно StratifiedKFold. StratifiedKFold е варијација на k-fold која враќа стратификувани делови: секој сет содржи приближно ист процент примероци од секоја целна класа како и целосниот сет[19].Како и во секоја друга крос-валидација,1 дел се користи за тестирање,а другите к-1 делови за тренирање.По испробување и експериментирање на повеќе различни вредности за к,на крајот најдобри резултати произведе к=8.На следната слика може да видиме како работи StratifiedKFold.



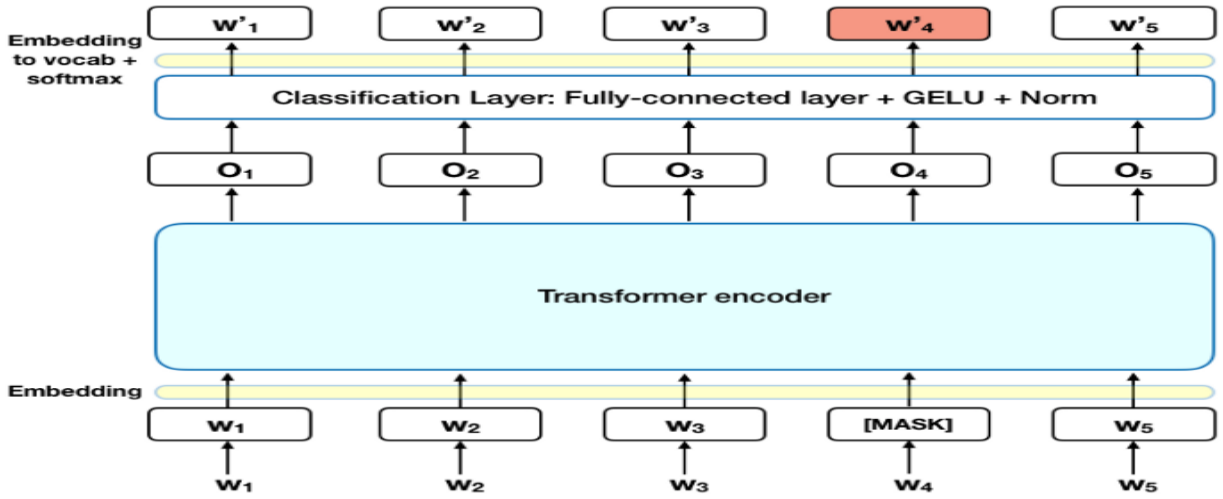
Слика 3. Приказ на StratifiedKFold. Извор: https://scikit-learn.org/stable/modules/cross_validation.html

3.5.Трансформери- BERT

BERT,што е кратенка за двонасочни енкодерски претстави од трансформатори, се базира на трансформатори, модел за длабоко учење во кој секој излезен елемент е поврзан со секој влезен елемент, а тежините меѓу нив динамички се пресметуваат врз основа на нивното поврзување. (Во НЛП, овој процес се нарекува внимание.) Историски гледано, јазичните модели можеа само последователно да читаат внесен текст -- или од лево кон десно или од десно кон лево -- но не можеа да ги прават и двете истовремено. БЕРТ е различен бидејќи е дизајниран да чита во двете насоки одеднаш. Оваа способност, овозможена со воведувањето на трансформаторите, е позната како двонасочност.Ова е една од причините поради која се решив да го употребам овој модел во моето истражување, за да ги проценам резултатите од при/ма на помодерните технологии.Имено,Гугл го вовел open-source BERT во 2018 година, додека пак пред тоа веќе се излезени некои од предтренираните модели.Неговиот пред-тренинг служи како основен слој на „знаење“ од кое треба да се изгради модел. Оттаму, BERT може да се прилагоди на постојано растечкото тело на содржини и прашања што може да се пребаруваат и да биде тунирање на спецификациите на корисникот. Овој процес е познат како трансфер на учење [20]. Трансформерот е дел од моделот кој на БЕРТ му дава зголемен капацитет за разбирање на контекстот и нејаснотијата во јазикот. Трансформерот го прави ова со обработка на кој било даден збор во однос на сите други зборови во реченицата, наместо да ги обработува еден по еден. Гледајќи ги сите околни зборови, трансформерот му овозможува на моделот BERT да го разбере целосниот контекст на зборот и затоа подобро да ја разбере намерата на учењето.

Ова се спротивставува на традиционалниот метод на обработка на јазикот, познат како word embedding, во кој претходните модели како GloVe и word2vec би го мапирале секој поединечен збор на вектор, кој претставува само една димензија.Овие модели побаруваат

голем корпус на лабелирани податоци.Од таа причина,тие не секогаш одговараат лесно на некои прашања од NLP. БЕРТ,пак користи метод на моделирање на маскиран јазик за да го задржи зборот во фокусот ,односно да има фиксно значење независно од неговиот контекст. БЕРТ потоа е принуден да го идентификува маскираниот збор само врз основа на контекстот. Во БЕРТ зборовите се дефинираат според нивната околина, а не според однапред фиксиран идентитет. Секој додаден збор го зголемува целокупното значење на зборот на кој се фокусира алгоритмот. Колку повеќе зборови се вкупно присутни во секоја реченица или фраза, толку зборот во фокус станува појасно. БЕРТ го зема предвид зголеменото значење со читање двонасочно, земајќи го предвид ефектот на сите други зборови во реченицата врз фокусниот збор.На следната слика,можеме да воочиме како работи овој модел.



Слика 4. Приказ на BERT моделот. Извор: https://towardsdatascience.com/bert-explained-state-of-the-art-language-model-for-nlp-f8b21a9b6270

На крајот,напоменувам декв во склоп на овој проект го користев BERT-base-uncased претрениранот модел.

3.6.Word embeddings

Финално,како последен дел од ова истражување,ќе се задржиме на 2 типа на word embeddings: Word2Vec и GloVe. Word2Vec е плитка, двослојна невронска мрежа која е обучена да реконструира лингвистички контексти на зборови.

Како свој влез зема голем корпус зборови и произведува векторски простор, типично од неколку стотици димензии, при што на секој единствен збор во корпусот му е доделен соодветен вектор во просторот.Векторите на зборовите се позиционирани во векторскиот простор така што зборовите што делат заеднички контекст во корпусот се наоѓаат во

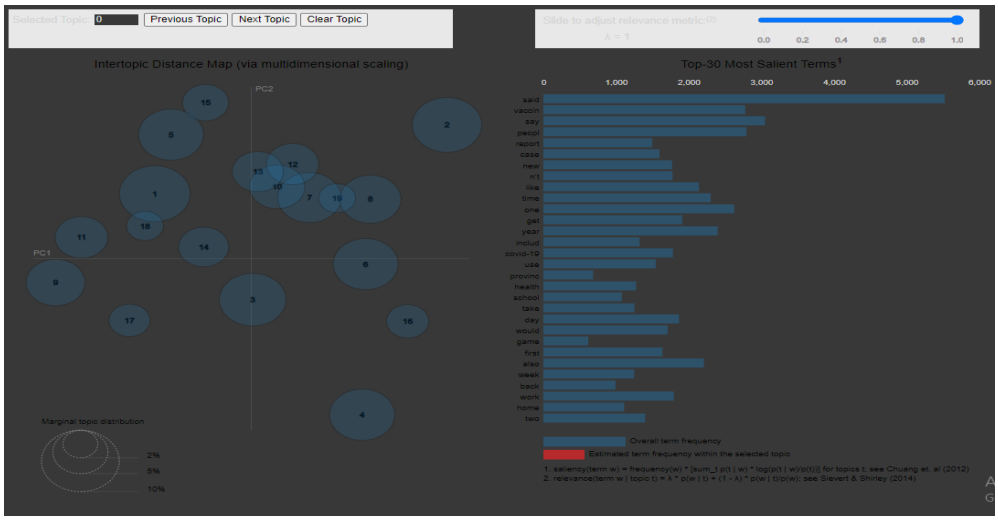
непосредна близина еден до друг во просторот. Word2Vec користи трик е познат во машинското учење.

Тој е едноставна невронска мрежа со еден скриен слој, и како и сите невронски мрежи, има тежини, а за време на тренингот, неговата цел е да ги прилагоди тие тежини за да ја намали функцијата на загуба. Сепак, Word2Vec нема да се користи за задачата за која е обучен, наместо тоа, ние само ќе ги земеме неговите скриени тежини и ќе ги користиме како word embeddings [21].

Од друга страна, GloVe е алгоритам за учење без надзор за добивање векторски претстави за зборови. Обуката се изведува на збирна глобална статистика за појавување на зборови од корпус, а добиените претстави прикажуваат интересни линеарни подструктури на векторскиот простор [22].Постојат повеќе верзии на пред-тренирани вектори од GloVe, во мојот случај јас искористив 2 сосема различни големини: мала верзија со 6B токени со 50 димензионални вектори и голема верзија со 840B токени со 300 димензионални вектори. Во текот на учењето,покрај овие модели,искористив и неколку основни за споредба,како што се MultinomialNB,BernoulliNB и SVC,секој со Count и TF-IDF векторизација.Што се однесува до Word2Vec и GloVe,имавме 6 модели: GloVe со 2 големини и Word2Vec, секој на 2 начини со Mean и TF-IDF embedding. Резултатите ќе ги погледнеме во следната секција.

4.Резултати

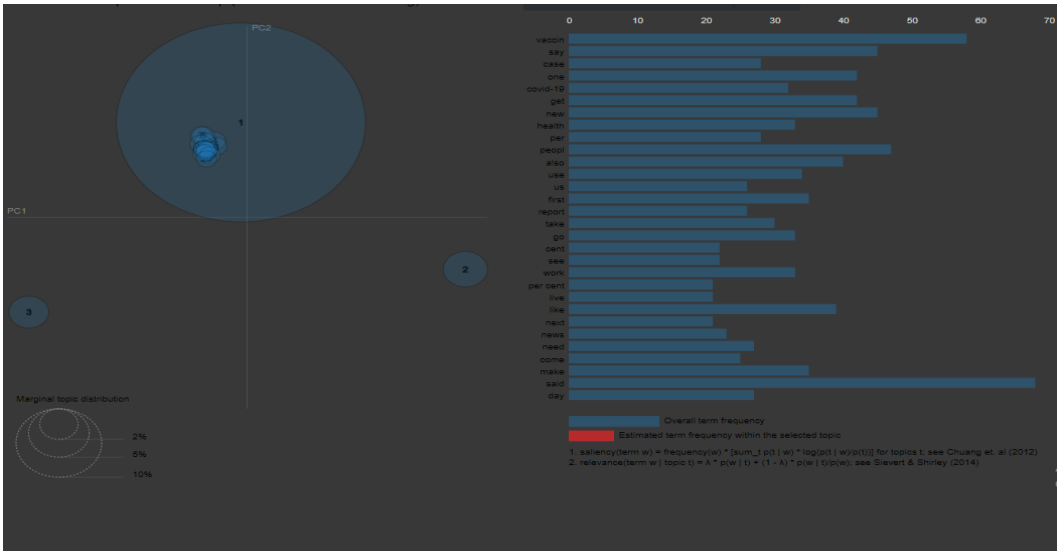
Кога сме завршиле со нашите методи за анализа на медиумската наклонетост,време е да направиме анализа на резултатите што сме ги добиле.Најпрвин ќе почнам со анализата на LDA моделите.Како што гледаме од сликата подолу,очигледно е дека LDA Gensim моделот со 19 теми(затоа што имаше најголема кохерентност на соодветниот график) добро си ја завршил работата затоа што дистрибуцијата на теми и зборови е прилично рамномерна. Тоа ни означува дека мерката за кохерентноста на темата дава добро тунирање на параметрите и значително ги подобрува перформансите на моделот.



Слика 5. Приказ на визуелизација на LDA Gensim моделот

За разлика од овој модел,можеме да видиме дека LDA моделот од SkLearn не ни дава добра распределба на темите и имаме големи кластери од кои повеќето се во 1 точка. Многу причини зошто моделот на тема со LDA не се снаоѓа добро:

1. Нема доволно податоци
2. Нема доволно разновидни извори на податоци
3. Не е направена темелна работа за чистење/преработка на самите текстуални податоци.
4. Параметарот min_df не е доволно добра опција за подесување/тунирање на моделот



Слика 6. Приказ на визуелизација на LDA SkLearn моделот

Дополнително, го користев додатокот TextBlob за да ни помогне да направиме брза анализа на сентимент и субјективност за секој документ во нашиот корпус. ние ќе генерираме резултат за сентимент кој се движи од -1 до 1. 1 е позитивно чувство, што значи дека статијата е главно позитивна, а -1 негативно чувство. Откако ќе го креираме нашиот резултат за чувства, би можеле да ги споредиме нашите теми низ нашите различни весници. Тоа не е точна наука, но може да ни даде увид.Покрај тоа, ќе креираме скор за субјективност кој се движи од 0 до 1, при што 0 е многу објективен стил на пишување,а 1 е многу субјективен стил на пишување. Оттука можеме да заклучиме дека документите што се поблиску до 1 може да бидат малку посубјективни и може да е присутна агенда во пишувањето.Од сликата подолу можеме да забележиме дека порталот Bridgemi ги добива највисоките оцени за субјективност особено на теми кои вклучуваат вакцини и КОВИД-19, а исто така се оценети со Bias==”Right”,па можеби е присутна некаква уредничка агенда.Од друга страна пак,порталот Channel News Asia е оценет како најобјективен.

	Topic	Sentimen	Source	Keywords	Subjectivity Score
0	11.0	0.096	Bridgemi	said, one, time, like, day, say, peopl, work, also, year	1.0
1	18.0	0.103	Bridgemi	said, say, peopl, year, like, get, time, nâ€™t, covid-19, one	1.0
2	2.0	0.071	Bridgemi	say, said, time, year, peopl, vaccin, first, like, also, work	1.0
3	10.0	0.039	Bridgemi	said, say, peopl, year, also, vaccin, one, make, like, servic	1.0
4	10.0	0.128	Bridgemi	said, say, peopl, year, also, vaccin, one, make, like, servic	0.95
5	11.0	0.113	Bridgemi	said, one, time, like, day, say, peopl, work, also, year	0.95
6	11.0	0.083	Bridgemi	said, one, time, like, day, say, peopl, work, also, year	0.944
7	2.0	0.096	Bridgemi	say, said, time, year, peopl, vaccin, first, like, also, work	0.941

Слика 7. Приказ на дел од табелата со оценки за субјективност

Од друга страна пак,Bridgemi воглавно добива оценка 0 за сентимент,а Channel News Asia добива поларизирачки оценки за сентимент(или високи или ниски),што значи дека користат чувства,но сепак оддржуваат објективност.

Следно,да направиме мала анлизаа на резултатите добиени со Scattertext методот. Таму пресметувавме прецизност на зборовите,нивната фрекфенција и скалираната Ф-мерка. На дел од следната табела може да ги видиме зборовите кои најчесто се појавуваат,односно тие кои имаат најголема фрекфенциј,групирани според нивната лабела за наклонетост.

term	Lean Left freq	Lean Right freq	Right freq	Center freq	dem_precision	dem_freq_pct	dem_hmean
.	160	17	97	63	0.583942	0.003358	0.006678
vaccine	132	4	54	17	0.694737	0.002770	0.005519
covid	129	9	337	21	0.271579	0.002707	0.005361
party	119	2	5	9	0.944444	0.002498	0.004982
one	115	23	123	69	0.440613	0.002414	0.004801
health	112	11	119	24	0.462810	0.002351	0.004677
people	111	21	109	101	0.460581	0.002330	0.004636
new	109	7	218	45	0.326347	0.002288	0.004543
paul	95	0	1	2	0.989583	0.001994	0.003980
would	94	8	64	42	0.566265	0.001973	0.003932

Слика 8. Приказ на дел од табелата со најфрекфентни зборови.

Може да забележиме дека кај левичарските медиуми главна тема се КОВИД-19, здравјето и вакцините,додека пак кај подесно ориентираните весници има термини како ‘people’, ‘new’,’one’.

За да се измерат перформансите на моделите со LinearSVC класификаторот потребна е метричка проценка – евалуација. Процесот на евалуација кажува колку добро моделот ќе генерализира на податоци надвор од нашиот примерок.

Првиот модел што го изработивме каде што како атрибути беа текстот и изворот има прецизност од 100.0,што повторно значи дека истиот нема доволно податоци за обработка и лесно оди во overfit и го отфрлив од понатамошна анализа.

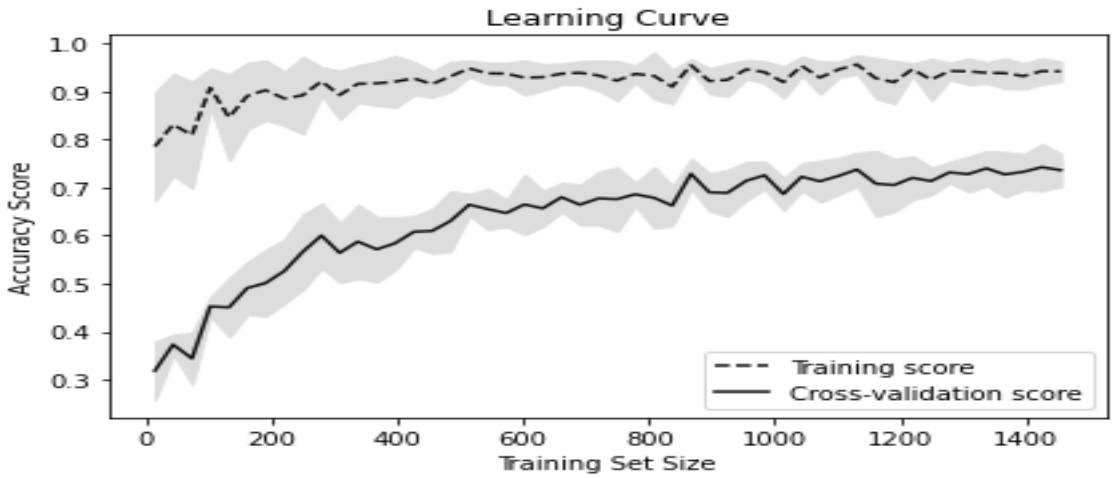
Другиот модел е сосема поинаква приказна.Кај него точноста изнесува 0.756,што за модел со 4 класи е сосема солидно.По извршената StratifiedKfold крос-валидација, може да забележиме мал пад во прецизноста,дадена во табелата,но сепак по експериментите к=8 одржуваше најдобро ниво на прецизност.

Fold number	Accuracy
1	0.78365
2	0.77884
3	0.76442
4	0.756
5	0.74134
6	0.75
7	0.73901
8	0.74144

Табела 1. Приказ на излезните вредности на StratifiedKFold.

Исто така,забележливо е дека прецизноста на крајот на тренинг множеството е 0.96, на валидациското множество 0.76,а на тест множеството 0.74.

Следно,на график ја нацртав кривата на учење на тренинг множеството,наспроти таа на StratifiedKFold за различни големини на тренинг множеството(од 1% до 100% на интервали од 5%).Очигледно е дека кривата на точноста кај крос валидацијата расте со земање на поголеми делови од податочното множество за тренинг,додека пак кај тренинг точноста таа по одредено време го достигнува врвот и го одржува истиот до крајот на тестирањето.



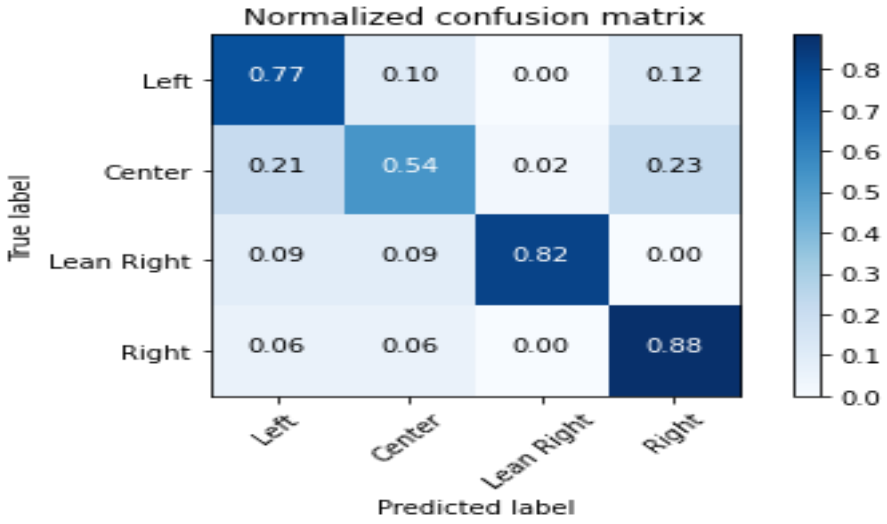
Слика 9. Приказ на графикот за learning curve.

Како последни мерки за евалуација ги имаме classification_report и матрица на конфузија. Од classification_report можеме да видиме дека кај сите мерки за точност,како што се прецизност,одзив и f1- мерката,класата 2 постојано постигнува најслаби резултати.Имено, тоа ни е класата со вредност за Bias=”Center”.Како што нагласив на почетокот,Center значи дека медиумот има релативна непристрастност,но не мора да значи дека е непристрасен, туку може да зничи дека “Center” може и да е мешавина од Left и Right bias, или пак може да испушти важни перспективи или да објавува индивидуални написи што прикажуваат пристрасност, додека често не прикажува многу предвидливи пристрасности.

classes	precision	recall	f1-score	support
1	0.75	0.77	0.76	106
2	0.69	0.54	0.60	84
3	0.90	0.82	0.86	22
4	0.77	0.88	0.82	121
accuracy			0.76	333
macro avg	0.78	0.75	0.76	333
weighted avg	0.75	0.76	0.75	333

Табела 2. Приказ на табелата со мерките за точност

Од матрицата за конфузија дека дел од лабелите за “Center” понекогаш се предвидуваат во “Left” или “Right” класите,најверојатно поради некои од горенаведените причини.



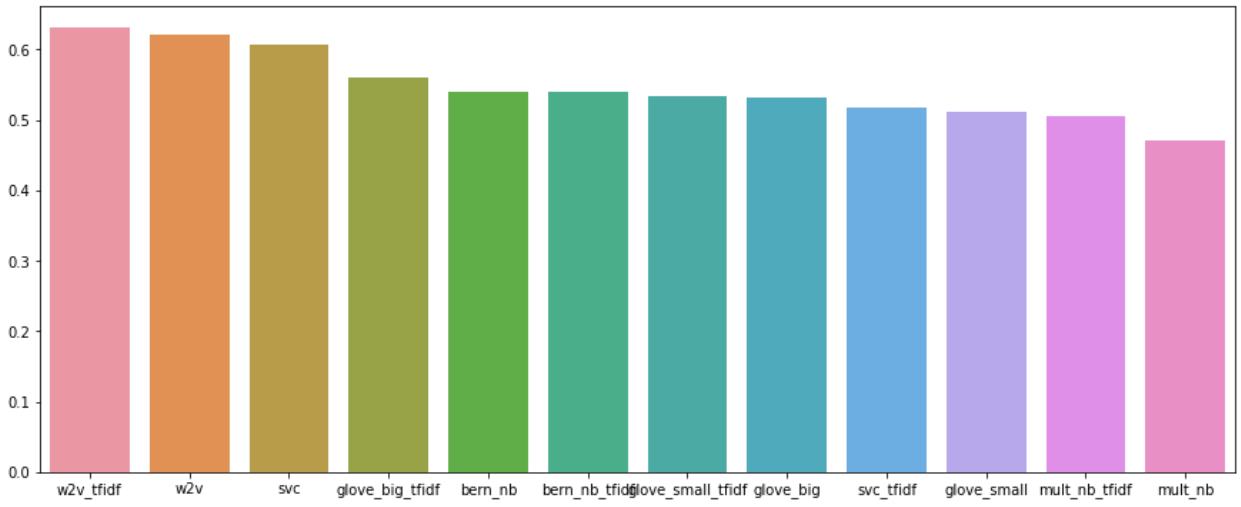
Слика 10. Приказ на нормализирана матрица на конфузија

BERT моделот беше трениран на 3 епохи, а потоа тестиран на тест множество,кое и/несуваше 10% од вкупниот датасет.Како мерка за евалуација повторно имаме classification_report,каде можеме да забележиме дека воглавно моделот дава добри ре/ултати,освен кај Lean Right класата,што го припишувам на многу малиот број на тест случаи,како и на инстанци во податочното множество.Сепак,вкупната прецизност од 0.83 е одлична за оваа количина на податоци и за BERT,воопшто.Целосниот поглед го имаме на следната слика.

classes	precision	recall	f1-score	support
Left	0.82	0.88	0.85	26
Center	0.70	0.76	0.73	21
Lean Right	0.00	0.00	0.00	5
Right	0.94	0.94	0.94	35
accuracy			0.83	87
macro avg	0.61	0.65	0.63	87
weighted avg	0.79	0.83	0.81	87

Табела 3. Приказ на табелата со мерките за точност кај BERT моделот

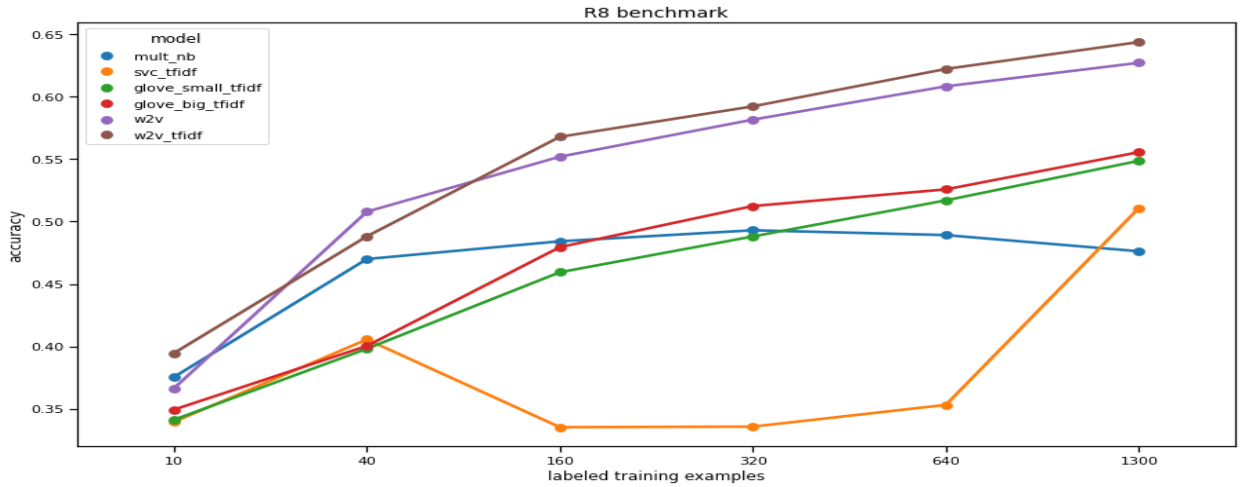
Крајно,ќе ги дискутираме перформансите кај моделите со word embeddings.По тренирањето на моделите и добивање на мерки за точност по преддефинираната 5-fold крос-валидација,се добиваат резултатите од графикот подолу.



Слика 11. Приказ на график со мерки за точност

Јасно воочливо е дека w2v моделите даваат најголема точност,како и тоа дека поголемата димензија на GloVe векторите повеќе помага при процесот на учење.Исто така,очигледно е дека TF-IDF векторизерот работи подобро на ова податочно множество од Mean.

Последната мерка за евалуација е валидација со Stratified ShuffleSplit крос-валидаторот. Тој обезбедува индекси за тренинг/тест за поделба на податоците во множества тренинг/тест множества.Овој метод за вкрстена валидација е спој на StratifiedKFold и ShuffleSplit, што враќа стратифицирани рандомизирани делови. Тие се прават со зачувување на процентот на примероци за секоја класа [23].Сепак,некои од деловите може да се поклопуваат.Ова е изведено за различни големини на тренинг множеството, а е видливо на сликата.



Слика 12. Приказ на график со развој на точноста на моделите

Се заклучува дека w2v моделите работат најдобро за мал број на податоци во тренинг множеството,како и тоа дека svc_tf-idf моделот поминува низ интересна варијација на намалување и зголемување на точноста.

Финално,ќе ги сумираме заклучоците во следната табела во која се прикажани сите употребени модели и пристапи кои спаѓаат во нагледувано учење и нивните резултати за точност на нашиот датасет во склоп на овој проект.Особено радува тоа што BERT моделот како најново достигнување на науката покажа најдобри резултати,што значи дека обработката на природните јазици,сепак се развива во вистинска насока.

Модел/пристап	Accuracy
BERT	0.8326
LinearSVC_Pipeline	0.7567
LinearSVC_Pipeline_Stratified 10-fold	0.7414
Word2Vec_TF-IDF	0.6305
Word2Vec	0.6212
SVC	0.6074
GloVe_big_TF-IDF	0.5606
BernoulliNB	0.5398
BernoulliNB_TF-IDF	0.5398
GloVe_small_TF-IDF	0.5341
GloVe_big	0.5323
SVC_TF-IDF	0.5173
GloVe_small	0.5116
MultinomialNB_TF-IDF	0.5046
MultinomialNB	0.4711

Табела 4. Финален приказ на мерките за точност на сите модели

4.1.Насоки за идни истражувања

Поради ограничени ресурси дел од идеите не беа реализирани.Во иднина би сакал да ги подобрам моделите за длабоко учење со додавање на уште повеќе податоци и понатамошно тунирање на параметрите за што е можно подобри перформанси.

Исто така,може да се истражат некои врски кои не беа доволно истражени,како на пример визуелизација на оценките за сентимент и субјективност,нивно користење во моделите за длабоко учење,како и подобро користење на вредностите за ARI индексот.

Најголеми подобрувања кои би можеле да се направат се користење на многу поголемо и поразновидно податочно множество,бидејќи е воочливо дека во најголемиот дел од проблематичните модели и визуелизации,тоа претставуваше најголем проблем.

5.Заклучок

Ова истражување разгледува повеќе различни пристапи кон анализирањето на медиумската наклонетост.Преку призмата на надгледувано и ненадгледувано учење се обидов да навлезам колку што умеев подлабоко во оваа тематика и мислам дека резултатот е прилично успешен.Медиумската пристрасност зафаќа голем дел од нашите животи,особено со растот на лажните вести кои полека почнуваат да ги надвладуваат реалните написи.

Сепак,чувствувам дека оваа тема на анализа на медиумска наклонетост сеуште не е доволно обработена во научниот свет,но сепак,науката на обработка на природните јазици се развива секојдневно и сигурен сум дека ќе гледаме се повеќе слични научни трудови.

Како и во моето претходно истражување за кластерирањето во областа на NLP, нагласувам дека LDA е многу специфичен модел којшто поседува многу параметри кои треба да се подесат,меѓутоа е релативно побрз од слични на нему модели,а и може да биде доста корисен ако се спои со модел за класификација на точност на кластерирање и евалуација,соодветно. Се обидов да применам модели од двата видови за да видам какви резултати би добил од двата пристапи,и се покажа дека иако различни,пристапите можат подеднакво добро да ги извршат своите задачи,како што се кластерирање,односно класификација.

Ненадгледуваното учење зема се поголем замав,како во областа на Машинското учење,така и во Обработката на природните јазици бидејќи кластерирањето може да ни ги прикаже колку некои објекти, односно зборови, се силно меѓусебно поврзани,а колку се раздалечени по значењето од други припадници на податочното множество.Надгледуваното учење пак, е повостановена пракса,меѓутоа и тоа масивно се развива со откривањето на нови модели и нови начини на процесирање на податоците.

За крај,ќе повторам дека би ме радувало,како мене,така и сите кои се занимаваат со оваа област, ако оваа тема продолжи да се развива со нови идеи за оптимизација и подобрување на резултатите.

6.Референци

[1]. Felix Hamborg, Norman Meuschke, and Bela Gipp. Bias-aware news analysis using matrix-based news aggregation. International Journal on Digital Libraries.

[2]. Zizi Papacharissi and Maria de Fatima Oliveira. 2008. News Frames Terrorism: A Comparative Analysis of Frames Employed in Terrorism Coverage in U.S. and U.K. Newspapers. The International Journal of Press/Politics, 13(1):52–74.

[3]. John McCarthy, Larissa Titarenko, Clark McPhail, Patrick Rafail, and Boguslaw Augustyn. 2008. Assessing stability in the patterns of selection bias in newspaper coverage of protest during the transition from communism in Belarus. Mobilization: An International Quarterly, 13(2):127–146.

[4]. Eric P.S. Baumer, Francesca Polletta, Nicole Pierski, and Geri K. Gay. 2017. A Simple Intervention to Reduce Framing Effects in Perceptions of Global Climate Change. Environmental Communication.

[5]. Doumit, Sarjoun & Minai, Ali. (2012). Online News Media Bias Analysis using an LDA-NLP Approach.

[6]. Laurenz A. Cornelissen, Lucia I. Daly, Qhama Sinandile, Heinrich de Lange, and Richard J. Barnett. 2019. A Computational Analysis of News Media Bias: A South African Case Study. In Proceedings of the South African Institute of Computer Scientists and Information Technologists 2019 (SAICSIT '19). Association for Computing Machinery, New York, NY, USA, Article 25, 1–10.

[7]. Mohit Iyyer, Peter Enns, Jordan L. Boyd-Graber, and Philip Resnik. 2014. Political ideology detection using recursive neural networks. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, Volume 1: Long Papers, pages 1113–1122.

[8]. Amr Ahmed and Eric P. Xing. 2010. Staying informed: supervised and semi-supervised multi-view topical analysis of ideological perspective. In Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP '10). Association for Computational Linguistics, USA, 1140–1150.

[9]. https://github.com/codelucas/newspaper

[10]. https://www.allsides.com/media-bias/media-bias-ratings

[11]. https://radimrehurek.com/gensim/models/phrases.html

[12]. https://towardsdatascience.com/topic-model-visualization-using-pyldavis-fecd7c18fbf6

[13]. https://towardsdatascience.com/evaluate-topic-model-in-python-latent-dirichlet-allocation-lda-7d57484bb5d0

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

[15]. https://github.com/JasonKessler/scattertext

[16]. https://pypi.org/project/textstat/

[17]. https://en.wikipedia.org/wiki/Automated_readability_index

[18]. https://scikit-learn.org/stable/modules/generated/sklearn.svm.LinearSVC.html

[19]. https://scikit-learn.org/stable/modules/cross_validation.html

[20]. https://searchenterpriseai.techtarget.com/definition/BERT-language-model

[21]. https://israelg99.github.io/2017-03-23-Word2Vec-Explained/

[22]. Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global Vectors for Word Representation.

[23]. https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.StratifiedShuffleSplit.html