REPUBLIKA E SHQIPËRISË
UNIVERSITETI POLITEKNIK I TIRANËS
FAKULTETI I TEKNOLOGJISË SË INFORMACIONIT
DEPARTAMENTI I INXHINIERISË INFORMATIKE

NELDA KOTE

PËR MARRJEN E GRADËS

"DOKTOR"

NË: "TEKNOLOGJITË E INFORMACIONIT DHE KOMUNIKIMIT"

DREJTIMI "INXHINIERI INFORMATIKE"

DISERTACION

SISTEMET INTELIGJENTE PËR PËRPUNIMIN E INFORMACIONIT: ALGORITME PËR PËRPUNIMIN E GJUHËS SHQIPE DHE GJETJEN E INFORMACIONIT NË TEKSTE NË SHQIP

Udhëheqës shkencor

Prof. Asoc. Marenglen Biba

TIRANË, 2021

SISTEMET INTELIGJENTE PËR PËRPUNIMIN E INFORMACIONIT: ALGORITME PËR PËRPUNIMIN E GJUHËS SHQIPE DHE GJETJEN E INFORMACIONIT NË TEKSTE NË SHQIP

Disertacioni

i paraqitur në Universitetin Politeknik të Tiranës

për marrjen e gradës

"Doktor"

në:

"Teknologjitë e Informacionit dhe Komunikimit"

Drejtimi "Inxhinieri Informatike"

Nga

Znj. Nelda Kote

TIRANË, 2021

JURIA PËR VLERËSIMIN E DISERTACIONIT PËR FITIMIN E GRADËS SHKENCORE “DOKTOR”

Miratuar

Me vendimin e Këshillit të Profesorëve të FTI-së Nr: 18, datë: 15.07.2021.

Kryetari i Jurisë:	Prof. Dr. Aleksandër XHUVANI
Anëtar i Jurisë (Oponent):	Prof. Dr. Elinda MEÇE
Anëtar i Jurisë:	Prof. Asoc. Igli TAFA
Anëtar i Jurisë (Oponent):	Akademik Neki FRASHËRI
Anëtar i Jurisë:	Prof. Asoc. Alda KIKA

Dekan i Fakultetit të Teknologjisë së Informacionit

Prof. Dr. Elinda MEÇE

TABELA E PËRMBAJTJES

ABSTRAKT

Përhapja e përdorimit të teknologjisë në dekadat e fundit pothuajse në çdo aspekt të jetës ka rezultuar në gjenerimin e një sasie të madhe të dhënash, sidomos të dhëna në formën e pastrukturuar siç është teksti. Rritja e sasive të të dhënave të krijuara, vit pas viti, sjell domosdoshmërinë e zhvillimit të metodave të reja dhe përmirësimin e atyre aktuale për përpunimin e të dhënave të pastrukturuara në formë tekst. Gjithë ky volum i madh të dhënash duhet të përpunohet dhe prej tij të nxirren informacione të vlefshme për t'u aplikuar në fusha të ndryshme.

Përpunimi i gjuhës natyrale dhe gjetja e informacionit nga të dhëna tekst janë fusha të Inteligjencës Artificiale kombinuar me Linguistikën Kompjuterike të cilat na japin mjetet e nevojshme për të përpunuar dhe për të nxjerrë informacion të vlefshëm nga sasi shumë të mëdha të dhënash në formë teksti. Dekadën e fundit, mjete të tilla kanë patur një zhvillim shumë të vrullshëm dhe përdorim shumë të gjerë. Aplikimet e tyre janë shumë të gjëra si në analizimin e ndjenjave, parapërpunimin linguistik dhe sintaksor, nxjerrjen e informacionit të kërkuar, sistemet e pyetje-përgjigje, sistemet tekst-zë, sistemet e dialogut, klasifikimin e teksti, etj. Nga realizimi i një kërkimi sistematik mbi punime shkencore në këto fusha për gjuhën shqipe kemi vënë re se ka disa punime modeste dhe mungesë në mjete me akses publik me saktësi të lartë për t'u përdorur. Kjo për arsye të ndryshme, si për nga kompleksiteti i gramatikës së gjuhës shqipe po dhe nga interesi i ulët i përdorimit të gjuhës shqipe në platforma apo motorë kërkimi të mëdhenj në rrjet duke qenë se përdoret nga një komunitet relativisht i vogël njerëzish kundrejt gjuhëve të tjera. Duke u nisur nga ky vëzhgim jemi motivuar për të zhvilluar mjete për përpunimin dhe gjetjen e informacionit në tekste në shqip.

Mediet sociale gjithmonë e më shumë po ndikojnë në mënyrën se si njerëzit, po komunikojmë dhe po shprehim ndjenjat dhe mendimet e tyre. Kjo ka rezultuar në një sasi të madhe të dhënash, si opinione të cilat duke i analizuar mund të përfitojmë informacion shumë të nevojshëm. Opinionet në mediet *online* kanë një ndikim të madh në jetën e përditshme dhe zhvillimi i mjeteve kompjuterike për analizimin e tyre dhe nxjerrjen e informacionit ka një rëndësi të veçantë. Analizime të tilla kanë rëndësi të veçantë sidomos në fushën e biznesit, ku mbi bazën e informacioneve të nxjerra nga analizimi i opinioneve në mediet *online* mund të merren vendime të rëndësishme për përmirësimin e produkteve apo shërbimeve për të rritur të ardhurat e biznesit. Opinion Mining, ose siç njihet ndryshe analizimi i ndjenjave, është fusha që merret me analizimin dhe nxjerrjen e informacionit nga opinione të shprehura në mediet *online*.

Në pjesën e parë të këtij disertacioni kemi vlerësuar eksperimentalisht performancën e algoritmeve të të mësuarit e automatizuar për detyrën e klasifikimit të opinioneve sipas polaritetit të ndjenjës së shprehur në to, në pozitive apo negative. Për të realizuar këtë vlerësim kemi ndërtuar një korpus të etiketuar me opinione të nxjerra nga medie *online* të njohura në gjuhën shqipe. Janë marrë në vlerësim 50 algoritme të implementuara në

platformën Weka dhe dy rrjete neurale të implementuara në platformat Keras dhe TensorFlow duke kombinuar karakteristika të ndryshme. Nga vlerësimi eksperimental ka rezultuar se rrjeti neural i thjeshtë, *bag-of-words*, ka performancë më të mirë se rrjeti CNN dhe 50 algoritmet e tjera. Duke marrë në konsideratë vetëm 50 algoritmet e implementuar në platformën Weka, algoritmet më performante janë Naive Bayes Multinomial dhe RBF Network, performanca e të cilave rritet me përdorimin e TF-IDF dhe n-gram me vlerë min=1 dhe max=2.

Një tjetër mjet shumë i rëndësishëm në fushën e përpunimit të gjuhës natyrore dhe që përdoret në shumë sisteme të tjera është etiketuesi morfologjik dhe temëzuesi. Etiketuesi morfologjik ka si funksion t'i përcaktojë çdo fjale në një fjali etiketën përkatëse të pjesëve të ligjëratës, kurse temëzuesi të gjejë temën e fjalës. Në mënyrë që të bëhet një analizë më e mirë e të dhënave të pastrukturuara, tekst, në shumë sisteme është e nevojshme përdorimi i etiketave të pjesëve të ligjëratës dhe temëzuesit.

Në pjesën e dytë të këtij disertacioni kemi ndërtuar një etiketues morfologjik, për pjesët e ligjëratës dhe karakteristikave morfologjike, dhe një temëzues për tekst në gjuhën shqipe. Duke qenë së në gjuhën shqipe ka mungesa dhe në korpuse të etiketuara për t'u përdorur në aplikime të ndryshme të përpunimit të gjuhës natyrore, kemi krijuar nga fillimi një korpus të etiketuar. Në këtë disertacion propozojmë një skemë etiketimi për gjuhën shqipe duke u bazuar në një skemë shumëgjuhëshe Universal Dependencies që është përdorur për etiketimin e një korpusi me 184,597 fjalë (23,686 fjali). Korpusi i etiketuar është përdorur për të trajnuar një model të bazuar në parserin Turku Neural Parser Pipeline. Në këtë disertacion kemi paraqitur etiketuesin e parë morfologjik, për pjesët e ligjëratës dhe karakteristikat morfologjike, dhe temëzuesin e parë për gjuhën shqipe me akses publik me performancë shumë të mirë në çdo fazë të tij. Ky etiketues dhe temëzues mund të shërbejë si pikënisje për të vazhduar punën e gjatë në realizimin e një etiketuesi dhe temëzuesi zyrtar për gjuhën shqipe.

Fjalë kyçe: *Sisteme Inteligjente, Përpunimi i Gjuhës Natyrale, Opinion Mining, Etiketimi i Pjesëve të Ligjëratës, Temëzim, Analizimi i Opinioneve, Analizimi i Ndjenjave, Gjetje Informacioni, Gjuha Shqipe.*

ABSTRACT

Technology is used in every aspect of life, and large volumes of data, especially unstructured data as text, are generated day by day. The increasing volumes of data lead to the necessity to develop new methods and improve the current ones for processing unstructured data as text. All these large volumes of data need to be processed, and the extracted information can be used in various fields.

Natural Language Processing and Information Retrieval are Artificial Intelligence subfields combined with Computer Linguistics that aim to develop tools to process and retrieve information from large volumes of text data. In the last decade, these tools are rapidly developed and widely used. They are widely used in emotion analysis, linguistic and syntactic preprocessing, information extraction, question-answer systems, text to voice systems, dialogue systems, text classification, etc. From a systematic review of the research works in these fields for the Albanian language, we can conclude that there are some modest works done, but there are no free-to-use tools with high accuracy. The complexity of the Albanian language and the low interest to develop tools for a language used by a relatively small community are the most important factors of this situation. This situation motivated us to develop preprocessing and information retrieval tools for text in the Albanian language.

Social media is increasingly influencing the way people are communicating and expressing their feelings and thoughts. A large volume of data, as opinions, are generated in social media, and by analyzing them we can obtain important information. The opinions expressed in social media have a great impact on daily life and developing applications to analyze them is particularly important. Business is one of the fields where analyzing customers' opinions about products or services can influence future decision-making to improve products or services and increase business income. Opinion Mining or sentiment analysis is the field that analyzes and extracts information from opinions expressed in online media.

Firstly in this dissertation, we have experimentally evaluated the performance of machine learning algorithms for opinions classification based on the polarity of the expressed sentiment, as positive and negative. We have created a text opinion annotated corpus by collecting opinions in the Albanian language from well-known online media in Albania. We evaluated the performance of 50 machine learning algorithms implemented in Weka, and two neural networks implemented in Keras and TensorFlow. The experimental results show that the simple bag-of-words model has better performance than the CNN model and the 50 other algorithms. Taking into consideration only the experimental results of the 50 algorithms, the best performing algorithms are Naive Bayes Multinomial and RBF Network. The performance of these algorithms is enhanced by using TF-IDF and n-grams with min = 1 and max = 2.

The morphological tagger and the lemmatizer are important tools of natural language processing used by a wide set of systems. The morphological tagger aims to define to each token in a sentence the corresponding part-of-speech tag, and the lemmatizer aims to find the lemma of each token in the sentence. In many systems, to make a better analysis of the text data, it is necessary to use these tools.

Secondly in this dissertation, we present a part-of speech and morphological tagger and a lemmatizer for text in the Albanian language. Due to the fact that in the Albanian language there is a lack of annotated corpora, we have created an Albanian part-of-speech corpus. We have proposed an annotation schema based on Universal Dependencies schema for part-of-speech and morphological characteristics annotation. The proposed schema is used to annotate a corpus of 184,597 tokens (23,686 sentences). The annotated corpus is used to train and evaluate a model of Turku Neural Parser Pipeline parser for the Albanian language. In this dissertation, we have presented the first morphological tagger, for part-of-speech and morphological characteristics annotation, and the first lemmatizer for the Albanian language under an open license with high performance in each of its stages. This tagger and lemmatizer can be a starting point to the development of the official morphological tagger and lemmatizer for the Albanian language.

Keywords: *Intelligent Systems, Natural Language Processing, Opinion Mining, Part-of-Speech Tagging, Lemmatizing, Opinion Analyzing, Sentiment Analysis, Information Extraction, Albanian Language.*

FALENDERIME

Në radhë të parë dua të shpreh mirënjohjen për udhëheqësin shkencor, Prof. Asoc. Marenglen Biba, i cili më drejtoi për të punuar në fusha me shumë interes siç janë përpunimi i gjuhës natyrale dhe nxjerrja e Informacionit dhe në veçanti në aplikimin e tyre për gjuhën shqipe. Mbështetja, ndjekja dhe diskutimet e vazhdueshme me të kanë qenë faktor i rëndësishëm në ecurinë dhe përmbylljen me sukses të këtij disertacioni, duke propozuar mjete të rëndësishme të përpunimit dhe nxjerrjes së informacionit të tekstit në gjuhën shqipe.

Falenderoj përzemërsisht Dekanin e Fakultetit të Teknologjisë së Informacioni, Prof. Dr. Elinda Meçe, për mbështetjen e vazhdueshme dhe ndjekjen e mbarëvajtjes së këtij cikli studimesh.

Gjithashtu, dua të falënderoj dhe të gjithë kolegët e mi dhe në veçanti Përgjegjësin e Departamentit të Bazave të Informatikës, Prof. Aleksandër Xhuvani, dhe Përgjegjësin e Departamentit të Inxhinierisë Informatike, Dr. Enida Sheme, për përkrahjen e vazhdueshme.

Një falënderim edhe për bashkëpunëtorët në Grupin NLP, në Universitetin e Turkut, Finlandë, falë bashkëpunimit me të cilët arritëm që të propozojmë një etiketues morfologjik dhe temëzues për gjuhën shqipe.

Në veçanti, dua të falenderoj familjen time për mbështetjen pakufi dhe motivimin gjatë gjithë këtij cikli studimi.

Nelda Kote

Tiranë, 2021

LISTA E FIGURAVE

LISTA E TABELAVE

LISTA E GRAFIKËVE

FJALOR TERMINOLOGJIK

Analizë morfologjike (angl. morphological parsing) – Analizimi i karakteristikave morfologjike të formës aktuale të fjalës në një fjali.

Analizë sintaksore (angl. syntactic parsing) – Analizimi i lidhjeve sintaksore të fjalëve në një fjali, në formë peme.

Analizimi i ndjenjave - angl. Sentiment Analysis (SA)

***Bag-of-words* -** teksti trajtohet si një bashkësi fjalës pa lidhje midis tyre.

Bashkëtrajnim – angl. co-training

Etiketim – procesi i shoqërimit të një etikete të caktuar pjesë e një grupi etiketash çdo entiteti në fjali ose çdo dokumenti.

Etiketimi i karakteristikave morfologjike – procesi i etiketimit të çdo fjale të një fjalie me një etiketë që përfaqëson karakteristikat morfologjike të formës aktuale të fjalës.

Etiketimi i pjesëve të ligjëratës – procesi i etiketimit të çdo fjale të një fjalie me një etiketë që përfaqëson pjesën e ligjëratës që ajo i përket.

Etiketues i pjesëve të ligjëratës (angl. Part-of-Speech Tagger (POS)) – një program i cili përcakton etiketën e pjesëve të ligjëratës të formës aktuale të fjalës në një fjali.

Etiketues morfologjik – një program i cili përcakton etiketën e pjesëve të ligjëratës dhe etiketat e karakteristikave morfologjike të formës aktuale të fjalës në një fjali.

Gjetja e Informacionit të Kërkuar (GJIK) - angl. Information Retrieval (IR)

Inteligjenca Artificiale (IA) – angl. Artificial Intelligence (AI)

Korpus – një bashkësi dokumentesh që përdoren për trajnimin dhe vlerësimin e një modeli.

Linguistika Kompjuterike (LK) - angl. Computational Linguistics (CL)

Nxjerrja e Informacionit të Strukturuar (NIS) - angl. Information Extraction (IE)

Opinion Mining (OM) – fushë studimi që merret me analizimin e opinioneve në mediet *online*.

Përmbledhja e Tekstit (PT) - angl. Text Summarization (TS)

Përpunimi i Gjuhës Natyrale (PGJN) - angl. Natural Language Processing (NLP)

Rrënjëzim (angl. stemming) – procesi i gjetjet së rrënjës së fjalës.

Rrënjëzues (angl. stemmer) – një program i cili përcakton rrënjën e fjalës.

Rrjet Neural Artificial (RNA) - angl. Artificial Neural Network (ANN)

Saktësia (angl. Accuracy) – njësi matëse për vlerësimin e cilësisë së një modeli.

Sistemet Pyetje-Përgjigje (SPP) – angl. Question-Answering (QA)

Sistemi i Dialogut në Gjuhën e Folur (SDGJF) – angl. Spoken Language Dialogue System (SLDS)

Temëzim (angl. lemmatize) – procesi i gjetjes së temës së fjalës.

Temëzues (angl. lemmatizer) – një program i cili përcakton temën e fjalës.

Të bazuara në fjalor gjuhësor - angl. dictionary-based

Të bazuara në korpus - angl. corpus-based

Të bazuara në leksik - angl. lexicon-based

Të mësuarit e pakontrolluar - angl. unsupervised learning

Të mësuarit e përforcuar - angl. reinforcement learning

Të mësuarit e kontrolluar - angl. supervised learning

Të mësuarit e thelluar – angl. Deep Learning (DL)

Të mësuarit gjysmë të kontrolluar - angl. semi-supervised learning

Të mësuarit e automatizuar (MA) – angl. Machine Learning (ML)

Vetëtrajnim – angl. self-training

KREU 1

HYRJE

1.1. Motivimi i studimit

Në dekadat e fundit sasia e informacionit të krijuar në mediet *online* ka qenë gjithmonë në rritje. Ndër tipet e të dhënave të krijuara janë dhe të dhënat e pastrukturuara, siç është teksti. Analizimi i të dhënave dhe gjetja e informacionit të dobishëm në to është shumë e rëndësishme në aspekte të ndryshme të jetës. Për këtë arsye të paturit e mjeteve sa më të mira është shumë e rëndësishme.

Përpunimi i Gjuhës Natyrale (PGJN) (angl. Natural Language Processing (NLP)) është një nga fushat e Inteligjencës Artificiale (IA) kombinuar me Linguistikën e cila na jep mjetet e nevojshme për të përpunuar sasi shumë të mëdha të dhënash në formë teksti. Vitet e fundit këto mjete kanë patur një zhvillim të madh duke arritur rezultate shumë të mira. Disa mjete PGJN që kanë një rëndësi të veçantë në përpunimin e tekstit janë parsimi, etiketimi morfologjik, temëzimi, rrënjëzimi, etj. Mjetet PGJN kanë aplikim të gjerë edhe në fusha të tjera të Inteligjencës Artificiale për të dhëna tekst. Aplikime që mund të përmendim janë klasifikimi i tekstit, analizimi i ndjenjave, sistemet e përkthimit automatik nga një gjuhë në tjetrën, gjetja e informacionit, sistemet e dialogut, sistemet tekst-zë, etj. Një aplikim më konkret i këtyre mjeteve për shembull është në sistemet Internet of Things (IoT), si Alexa apo Sirio që i përdorin ato për përpunimin paraprak të tekstit.

Opinion Mining (OM), ose analizimi i ndjenjave (angl. Sentiment Analysis (SA)) është një tjetër fushë me interes në Inteligjencën Artificiale (IA) që ka për qëllim gjetjen e informacionit nga opinione të shprehura në mediet *online*. Në raportin e vitit 2020 të BrightLocal (2021) të "Anketës Lokale të Opinioneve të Konsumatorit" përcaktohet se 87% e konsumatorëve lexojnë opinionet në mediet *online*. Secili nga ne përpara se të rezervojë një hotel, lexon komentet që kanë shkruar vizitorët e atij hoteli në medie të ndryshme *online*. Opinionet në mediet *online* kanë një ndikimi të madh në jetën e përditshme dhe zhvillimi i mjeteve kompjuterike për analizimin e tyre dhe nxjerrjen e informacionit nga to ka një rëndësi të veçantë. Analizime të tilla kanë rëndësi të veçantë sidomos në fushën e biznesit. Duke analizuar opinionet e klientëve në mediet *online* menaxherët mund të gjejnë informacione të rëndësishme për përmirësimin e cilësisë së produkteve apo shërbimeve dhe rrjedhimisht për të rritur dhe të ardhurat e biznesit.

Sisteme të tilla janë të varura nga gjuha dhe duhet që për çdo gjuhë të krijohen mjete specifike për atë gjuhë. Nga një kërkim sistematik për punime shkencore në ekzistencën e mjeteve PGJN për përpunimin e tekstit në shqip apo mjeteve për Opinion Mining kemi vënë re se janë zhvilluar vetëm disa punime modeste dhe nuk kemi ndonjë mjet me akses

publik me saktësi të lartë për t'u përdorur. Gjithashtu mungesa ka dhe në korpuse të etiketuara për qëllime të PGJN-së. Ky ishte një nga motivet që ndikoi në drejtimin e punës së këtij disertacioni në krijimin e mjeteve për përpunimin e teksteve në gjuhën shqipe dhe gjetjen e informacionit nga tekste në gjuhën shqipe.

1.2. Objektivat e studimit

Kontributi i këtij studimi është realizimi i sistemeve inteligjente për gjetjen e informacionit në tekste në gjuhën shqipe, konkretisht në analizimin e polaritetit të ndjenjës së shprehur në një opinion në mediet *online* dhe në krijimin e një sistemi etiketimi morfologjik dhe temëzimi të tekstit në gjuhën shqipe. Objektivat kryesore të këtij studimi janë:

- Vlerësimi i situatës aktuale të zhvillimit të teknikave për Opinion Mining dhe më në veçanti për klasifikimin e opinioneve sipas polaritetit të ndjenjës së shprehur në to.
- Vlerësimi i situatës aktuale të zhvillimit të teknikave për zhvillimin e etiketuesve morfologjikë dhe temëzuesve.
- Vlerësimi i situatës aktuale të zhvillimit të teknikave të përmendura më lart për gjuhën shqipe.
- Studimi i literaturës për morfologjinë dhe sintaksën e gjuhës shqipe.
- Krijimi i një korpusi opinionesh në gjuhën shqipe të etiketuara si opinione që shprehin një ndjenjë pozitive dhe negative.
- Vlerësimi i performancës së 50 algoritmeve MA dhe dy rrjeteve neurale në detyrën e klasifikimit të opinioneve sipas ndjenjës së shprehur në dy klasa si pozitive dhe negative.
- Propozimi i një skeme etiketimi morfologjik për gjuhën shqipe duke u bazuar në skemën shumëgjuhëshe Universal Dependencies (UD).
- Krijimi i një korpusi në gjuhën shqipe të etiketuar me etiketa për pjesët e ligjëratës dhe karakteristikat morfologjike dhe temën e fjalës.
- Implementimi i një etiketuesi morfologjik dhe temëzuesi për gjuhën shqipe duke përdorur korpusin e etiketuar nga fillimi dhe parserin Turku Neural Parser Pipeline.
- Diskutimi i zgjidhjeve të propozuara në këtë disertacion dhe i rezultateve eksperimentale.

1.3. Struktura e disertacionit

Struktura e disertacionit është si në vijim:

Kapitulli 2 përshkruan në terma të përgjithshme konceptin e PGJN-së si një kombinim i dy fushave kryesore atë të IA-së dhe LK-së. Më tej diskutohet për të mësuarin e automatizuar si një nga bazat e zhvillimit të aplikacioneve në fushën e PNGJ-së.

Kapitulli 3 trajton në mënyrë të detajuar Opinion Mining dhe në veçanti detyrën e klasifikimit të opinioneve. Në këtë kapitull diskutohen teknikat e aplikuara deri më sot në zhvillimin e aplikacioneve që klasifikojnë opinionet sipas polaritetit të ndjenjës së shprehur në to.

Kapitulli 4 trajton karakteristikat kryesore të gjuhës shqipe të cilat janë shumë të rëndësishme në ndërtimin e aplikacioneve që përpunojnë gjuhën e shkruar dhe të folur. Jemi ndalur në përshkrimin e elementëve morfologjikë dhe sintaksorë që janë të nevojshëm për realizimin e modeleve të klasifikimit të opinioneve, etiketuesin gramatikor dhe temëzuesin.

Kapitulli 5 përshkruan në terma të përgjithshme problemin e etiketimit morfologjik, temëzimit, rrënjëzimit dhe parsimit. Më tej diskutohen teknikat e përdorura në këta fusha dhe në pjesën e fundit diskutohet me detaje arkitektura e parserit Turku Neural Parser Pipeline dhe skema e etiketimit Universal Dependencies (UD), që janë përdorur për të implementuar etiketuesin morfologjik dhe temëzuesin për gjuhën shqipe.

Kapitulli 6 paraqet në mënyrë të detajuar modelet e propozuara në këtë disertacion për klasifikimin e opinioneve, etiketimin morfologjik dhe temëzimin në gjuhën shqipe. Në pjesën e parë të këtij kapitulli paraqitet korpusi dhe modelet e implementuara për vlerësimin e performancës së 50 algoritmave MA dhe dy rrjeteve neurale në detyrën e klasifikimit të opinioneve sipas polaritetit të ndjenjës së shprehur në to. Në pjesën e dytë diskutohet implementimi i etiketuesit morfologjik dhe temëzuesit, duke u nisur nga skema e etiketimit të përdorur, mënyra se si është krijuar korpusi i etiketuar dhe hapat e ndjekur për trajnimin dhe vlerësimin e modelit të etiketuesit morfologjik dhe temëzuesit.

Kapitulli 7 analizon në mënyrë të detajuar rezultatet e vlerësimeve eksperimentale dhe përfundimet e arritura gjatë punës së këtij disertacioni. Gjithashtu paraqiten edhe përmirësime të mundshme dhe drejtime për punime në të ardhmen në të njëjtën fushë.

KREU 2

PËRPUNIMI I GJUHËS NATYRALE

2.1. Hyrje

Përpunimi i Gjuhës Natyrale (PGJN) është një fushë shumë disiplinore e Inteligjencës Artificiale (IA) që ka si qëllim krijimin e mjeteve kompjuterike për përpunimin e gjuhës natyrale. Konsiderohet si aplikimi inxhinierik i fushës së Linguistikës Kompjuterike (LK) (angl. Computational Linguistics (CL)). PGJN-ja kombinon shkencën kompjuterike me gjuhësinë, matematikën, psikologjinë, biologjinë, etj., për të krijuar mjete për përpunimin e gjuhës natyrale. Ajo ka si qëllim të analizojë, kuptojë, nxjerrë informacion dhe të krijojë përmbajtje në gjuhën e përdorur nga njerëzit, si tekst ashtu dhe zë. Në vitet e fundit PGJN-ja ka patur një zhvillim të vrullshëm dhe është një nga fushat më të rëndësishme, duke kaluar nga sisteme të thjeshta të bazuara në modele teorike në sisteme më të qëndrueshme dhe më të sakta të bazuara në të mësuarin nga korpuse të mëdha të dhënash (Clark, Fox, & Lappin, 2010). Disa prej aplikimeve më të rëndësishme të viteve të fundit të PGJN-së janë diskutuar në çështje 2.3 .

Qëllimet kryesore pse kompjuterët duhet të përpunojnë gjuhën natyrore janë: që të komunikojnë me njerëzit, që të mësojnë dhe të bëjnë përparime në shkencën e të kuptuarit dhe të përdorimit të gjuhës natyrale duke përdorur mjete të IA-së me linguistikën, psikologjinë konitive dhe neuroshkencën (Russell & Norvig, 2020).

Procesi i analizimit të gjuhës natyrale mund të konsiderohet si një detyrë e përbërë nga disa faza të cilat pasqyrojnë dallimet gjuhësore midis sintaksës, semantikës dhe pragmatikës. Dale (2010) ka përcaktuar se analiza e përpunimit të gjuhës natyrale kalon në pesë faza siç tregohet në Figurën 2.1.

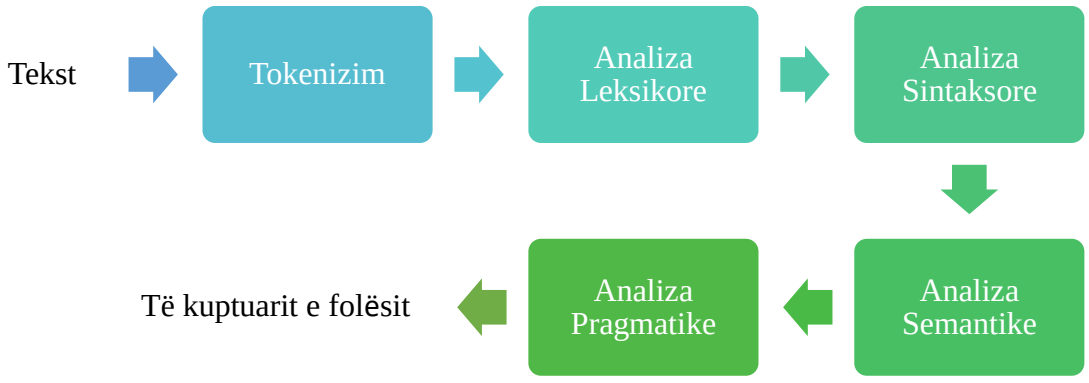


Figura 2.1 Fazat e analizimit në përpunimin e gjuhës natyrale (Dale, 2010)

Faza e parë tokenizimi dhe segmentimi i fjalive është një fazë shumë e rëndësishme dhe kritike duke qenë se tekstet zakonisht nuk përbëhen nga fjali të shkurtra, të mirë strukturuara dhe përcaktuara. Kjo fazë është më tepër e rëndësishme në gjuhët në të cilat segmentimi i fjalive nuk bazohet në ndarje sipas hapësirave. Në rastin e gjuhës shqipe përjashtim në tokenizim sipas hapësirave përbëjnë vetëm format e bashkuara të shkurtra të përemrave vetor. Faza e dytë analiza leksikore përcakton kuptimin leksikor të fjalës duke analizuar strukturën e fjalës. Faza e tretë analiza sintaksore ka si detyrë të analizojë strukturën sintaksore të një fjalie duke përcaktuar rendin dhe strukturën e fjalëve në fjali. Faza e katërt analiza semantike përcakton lidhjet midis fjalëve. Vendosja e kësaj faze pas fazës së analizës sintaksore mundëson një analizë më të mirë semantike duke u bazuar në informacionet e fazës së analizës sintaksore mbi strukturën e fjalisë. Faza e pestë dhe e fundit është analiza pragmatike e cila përcakton kuptimin dhe kontekstin e përdorimit të fjalës në fjali.

PGJN duhet të marrë në konsideratë çdo pjesë të strukturës së gjuhës. Në mënyrë që të kuptohet saktë teksti në gjuhën natyrale është shumë e rëndësishme që të kuptohet se çfarë përfaqësojnë dhe cilat janë ndryshimet midis niveleve të ndryshme të gjuhës. Nivelet e analizës linguistike janë: fonologjia, morfologjia, leksiku, sintaksa, semantika, ligjërata dhe pragmatika (Liddy, 1996).

Një nga problemet më të mëdha në sistemet PGJN është dykuptimshmëria. Kjo sepse në shumicën e gjuhëve një fjalë mund të përdoret në kuptime dhe funksione të ndryshme. Këto sisteme duhet të jenë të afta të identifikojnë se në kë kuptim dhe funksion është përdorur një fjalë e caktuar. Në këtë kontekst një sistem PGJN duhet të jetë në gjendje të marrë një vendim të drejtë në lidhje me kuptimin e fjalës, kategorinë e fjalës, strukturën sintaksore dhe qëllimin semantik të fjalës. Në të njëjtën kohë është e rëndësishme të maksimizohet cilësia e sistemit dhe të minimizohen gabimet në dykuptimshmëri. Në sistemet e para të zhvilluara të bazuara në rregulla të lidhura me një gjuhë të caktuar duhet të përcaktohen rregulla dhe procedura manualisht për të trajtuar sa më mirë dykuptimshmërinë. Në shumicën e rasteve këto rregulla janë të vështira për t'u përcaktuar dhe me saktësi të ulët. Kurse në sistemet të cilat bazohen në metoda të të mësuarit e automatizuar dhe që mësojnë nga korpuse shumë të mëdha, problemi i dykuptimshmërisë zgjidhet duke mësuar në mënyrë automatike kuptimin leksikor, strukturën e përdorimit të një fjale dhe lidhjet midis fjalëve nga korpusi. Metodat statistikore të përpunimit të gjuhës natyrale ofrojnë zgjidhje më të mira për adresimin e problemeve të dykuptimshmërisë (Manning & Schutze, 2000).

Mjetet e zhvilluara vitet e fundit, të cilat janë dhe shumë të sakta, bazohen në metoda të të mësuarit e automatizuar, si: rrjete neurale dhe të mësuarit e thelluar (angl. Deep Learning (DL)). Këto teknika nxjerrin në mënyrë automatike njohuri linguistike nga korpuse shumë të mëdha në vend që të kërkojnë zhvillimin e një sistemi në të cilin duhen përcaktuar manualisht rregullat për njohuritë linguistike të nevojshme për përpunimin e tekstit në gjuhë natyrale. Por nga ana tjetër, këto sisteme automatike të mësimit të njohurive

linguistike kanë nevojë për korpuse të mëdha dhe të etiketuara sipas detyrës që do të realizohet (Brill & Mooney, 1997).

2.2. Të mësuarit e automatizuar

Ideja e të mësuarit e automatizuar (MA) (angl. Machine Learning (ML)) është të lërë kompjuterin të veprojë në situata të ndryshme pa u programuar në mënyrë eksplicite, por duke adaptuar sjelljen mbi bazën e njohurive që ka mësuar nga të dhëna të mëparshme. Në MA kompjuteri studion të dhëna, mëson prej tyre dhe ndërton një model të cilin e përdor për të parashikuar veprime që duhet të ndërmarrë për probleme të ndryshme që duhet të zgjidh në të ardhmen (Russell & Norvig, 2020).

Përparësia e përdorimit të MA dhe jo e programimit eksplicit për zgjidhjen e një problemi të caktuar është se MA jep zgjidhje edhe për probleme që nuk janë të parashikueshme nga programuesi. Në praktikë është e pamundur që grupi zhvillues gjatë projektimit dhe implementimit të një programi të parashikojë të gjitha situatat e mundshme që do të duhet të zgjidhen nga programi në të ardhmen. Teknikat MA duke qenë se bazohen në të mësuarin nga të dhënat japin zgjidhje dhe në raste të paparashikuara në të ardhmen. Në shumë raste mund të ndodhë që dhe vetë programuesi të mos jetë në gjendje të gjejë zgjidhje për një problem të caktuar. Me përdorimin e një teknike MA, sistemi mund të jetë në gjendje të parashikojë veprimin e duhur për t'u ndërmarrë në këto raste. Kështu që komponentët e një programi mund të përmirësohen duke përdorur teknika MA (Russell & Norvig, 2020).

Në dekadat e fundit MA ka patur zhvillim dhe aplikim të gjerë në shumë fusha të jetës. Këtu mund të përmendim sisteme, si: motorët e kërkimit, makinat inteligjente me vetëdrejtim, sistemet e përpunimit dhe nxjerrjes së informacionit nga sasi të mëdha të dhënash, sistemet e pyetje-përgjigje, sistemet e dialogut njeri-kompjuter, sistemet e etiketimit morfologjik, sistemet e analizimit të ndjenjave, etj. Shumëkush nga ne në jetën e përditshme ka në përdorim sisteme të bazuara në MA. Për shembull, asistentët teknikë në të shkruar ne i kemi çdo ditë në përdorim në aplikacionet *chat*, duke na sugjeruar fjalën që mund të duam të shkruajmë në një mesazh mbi bazën e karaktereve të para. Në përpunimin e gjuhës natyrale, MA është një komponent kryesor për të modeluar dhe analizuar të dhënat tekst. Në të mësuarin e automatizuar aplikohen metoda të ndryshme për të realizuar detyra nga më të ndryshme.

Russell dhe Norvig (2020) përcaktojnë se ka tri tipologji të mësuari: të mësuarit e kontrolluar (angl. supervised learning), të mësuarit e pakontrolluar (angl. unsupervised learning) dhe të mësuarit e përforcuar (angl. reinforcement learning). Këto tipologji të mësuari diskutohen në vazhdim.

2.2.1. Të mësuarit e kontrolluar

Të mësuarit e kontrolluar (angl. supervised learning) është tipi kryesor i të mësuarit në MA dhe ka si qëllim të identifikojë lidhje midis një çifti të dhënash hyrje-dalje, pra të mësojë prej tyre mënyrën se si lidhen të dhënat e hyrjes me ato të daljes, dhe të ndërtojë një model i cili përdoret për të parashikuar të dhënat e daljes për një set të dhënash të reja të hyrjes. Procesi i të mësuarit është i kontrolluar sepse kur modelit i japim të dhëna të reja hyrje, ai do t'i krahasojë këto të dhëna me të dhënat nga të cilat ka mësuar dhe do të parashikojë daljen e mundshme. Bashkësisë së çifteve të dhëna hyrje-dalje ne i referohemi si të dhëna të etiketuara, ku hyrja është e dhëna dhe dalja është etiketa përkatëse. Bashkësisë së këtyre të dhënave të etiketuara ne i referohemi si korpus të dhënash të etiketuara. Në varësi të tipit të variablit në dalje, modeli mund të jetë regresion ose klasifikim. Modeli është regresion nëse parashikon një seri të vazhdueshme variablash në dalje. Modeli është klasifikues nëse parashikon daljen si një nga etiketat e një klase të përcaktuar etiketash (Russell & Norvig, 2020).

Për të përdorur metodat e të mësuarit të kontrolluar në një sistem të caktuar ne kemi nevojë për një korpus të etiketuar relativisht të madh. Ky është dhe disavantazhi më i madh i këtyre metodave, sepse etiketimi kërkon kohë dhe aftësi të mira në etiketim duke qenë se realizohet manualisht nga njerëzit. Gjithashtu, cilësia e etiketimit ka një ndikim mjaft të madh në performancën e këtyre teknikave. Sa më shumë të dhëna të etiketuara të përdoren për trajnimin e modelit, aq më shumë rritet performanca e tij.

Njësia e vlerësimit të një modeli është saktësia (angl. accuracy), numri i parashikimeve të sakta nga modeli për të dhëna të paetiketuara në hyrje nga të gjitha parashikimet e realizuara. Që të realizohet një vlerësim real i modelit duhet që korpusi i etiketuar të ndahet në dy ose tri korpuse: në korpusin e trajnimit dhe testimit ose në korpusin e trajnimit, optimizimit dhe testimit. Korpusi i trajnimit përdoret për të trajnuar modelin, pra që modeli të mësojë. Korpusi i testimit përdoret për të vlerësuar saktësinë e parashikimit të modelit, duke i dhënë modelit të trajnuar në hyrje të dhëna të paetiketuara, ai duhet të parashikojë etiketën dhe më pas të krahasohet etiketa e parashikuar me atë reale për të vlerësuar saktësinë e etiketimit. Korpusi i optimizimit përdoret për vlerësim gjatë trajnimit dhe nga ku modeli mund të mësojë të dhëna shtesë. Si korpusi i trajnimit dhe ai i testimit duhet të kenë të njëjtën formë, por duhet të përmbajnë të dhëna që nuk kanë lidhjen me njëra-tjetrën. Për shembull, nëse do të trajnojmë një model për të parashikuar nëse një dokument tekst i përket fushës së historisë apo jo, pjesë të tekstit që kanë lidhje me një dokument të caktuar duhet të ndodhen vetëm në një nga korpuset.

Në këtë disertacion kemi përdorur algoritme të ndryshme të bazuara në të mësuarin e kontrolluar për të ndërtuar modele që parashikojnë polaritetin e ndjenjave të opinioneve në shqip në dy klasa, pozitive dhe negative dhe rrjete neurale për të ndërtuar një etiketues morfologjik dhe temëzues për gjuhën shqipe.

2.2.2. Të mësuarit e pakontrolluar

Në ndryshim nga të mësuarit e kontrolluar në të mësuarin e pakontrolluar (angl. unsupervised learning) modeli krijohet duke mësuar nga të dhëna të paetiketuara. Metodat e të mësuarit e pakontrolluar përdoren në rastet kur është e pamundur të realizohet etiketimi manual i të dhënave për të krijuar korpuse të dhënash të etiketuara.

Një nga detyrat e të mësuarit të pakontrolluar është grupimi (angl. clustering). Grupimi ka si qëllim të identifikojë grupe të mundshme të dhënash në të dhënat e hyrjes. Grupi që i përket një të dhëne përcaktohet mbi bazën e ngjashmërisë së të dhënave midis tyre. Nuk kemi grupe paraprakisht të përcaktuara për të grupuar të dhënat dhe prandaj quhet të mësuarit e pakontrolluar, sepse grupet identifikohen gjatë analizimit të të dhënave. Të dhënat projektohen si pika në hapësirë dhe klasa që i përket një të dhëne përcaktohet nga afërsia me të dhënat e tjera. Metodat e të mësuarit të pakontrolluar mund të përdoren për përpunimin paraprak të të dhënave që përdoren në metoda të të mësuarit të kontrolluar (Russell & Norvig, 2020).

Avantazhi i këtyre metodave është që nuk kërkojnë korpuse të etiketuara që kanë kosto dhe vështirësi në etiketim.

2.2.3. Të mësuarit e përforcuar

Në të mësuarin e përforcuar (angl. reinforcement learning) modeli mëson nga një seri përforcimesh të formës: shpërblim dhe dështim. Një agjent marrje vendimesh mëson nga një seri shpërblimesh, që përcaktojnë se si është sjellja e tij, dhe duhet që të optimizojë numrin e shpërblimeve në të ardhmen. Për të realizuar këtë detyrë, agjenti mund të mësojë vlerën e një funksioni, një funksion Q, një rregull, etj. Në ndryshim nga të mësuarit e kontrolluar, në të mësuarin e përforcuar nuk është e nevojshme të kemi një korpus të etiketuar për të mësuar (Russell & Norvig, 2020).

2.2.4. Rrjeti neural artificial dhe të mësuarit e thellë

Vitet e fundit, Rrjeti Neural Artificial (RNA) (angl. Artificial Neural Network (ANN)) ka tërhequr vëmendjen e shumë kërkuesve për t'u përdorur për përpunimin e gjuhës natyrale. RNA ka për qëllim të nxjerrë karakteristikat duke kombinuar në mënyrë lineare të dhënat hyrëse dhe më pas të modelojë daljen si një funksion jo linear të këtyre karakteristikave. Ato paraqiten si një diagram rrjeti që përfshin nyje të lidhura midis tyre. Nyjet janë të vendosura në shtresa dhe arkitektura tipike e një rrjeti neural përmban tri shtresa: shtresa e hyrjes, shtresa e daljes dhe shtresa e fshehur (Moraes, et al., 2013).

Bazuar në tipologjinë e rrjetit, rrjetet neurale mund të kategorizohen në dy grupe, në *feedforward* dhe *recurrent*. Por, mund të kemi dhe rrjete që për nga tipologjia mund të jenë kombinime të këtyre dy grupeve. Një tip i veçantë i rrjeteve *feedforward* janë rrjetet Convolutional Neural Network (CNN). Këto rrjete janë përdorur fillimisht për përpunimin e imazheve dhe më pas janë adaptuar dhe në përpunimin e tekstit. Në rrjetet Recurrent

Neural Network (RNN) lidhjet midis neuroneve krijojnë një cikël të drejtuar të cilat japin mundësinë e përpunimit të sekuencave të informacionit ose tekstit (Zhang, et al., 2018).

Të mësuarit e thelluar (angl, Deep learning (DL)) është aplikimi i rrjeteve neurale për të mësuar dhe parashikuar duke përdorur rrjete me shumë shtresa. Avantazhi i tyre është aftësia më e lartë për të mësuar në krahasim me rrjetet neurale që zakonisht kanë një deri në tri shtresa dhe një sasi të vogël të dhënash. Long short-term memory (LSTM) është një tip special i rrjeteve RNN që ka aftësinë të mësojë në varësi afatgjate (Zhang, et al., 2018).

2.3. Aplikime të përpunimit të gjuhës natyrale

PGJN-ja ka një fushë aplikimi shumë të gjerë në analizimin e të dhënave tekst. Disa nga fushat dhe aplikacionet më të përhapura që përdorin mjete PGJN janë:

Nxjerrja e Informacionit të Strukturuar (NIS) (angl. Information Extraction (IE))

Qëllimi i NIS-së është analizimi i tekstit të pastrukturuar dhe nxjerrja e informacioneve të nevojshme të cilat ruhen në një formë të strukturuar, p.sh në një bazë të dhënash që mund të përdoret për qëllime të ndryshme si kërkim. Disa nga detyrat në NIS janë: nxjerrja e lidhjeve për gjetjen dhe klasifikimin e marrëdhënieve semantike midis entiteteve në tekst, grafi i njohurive për paraqitjen e njohurive nëpërmjet një grafi që ruan marrëdhëniet, nxjerrja e ngjarjeve për të identifikuar ngjarjet ku marrin pjesë entitetet, shprehjet e kohës për të nxjerrë kohën kur ka ndodhur një ngjarje dhe plotësimi i modelit për të identifikuar ngjarje apo situata që përsëriten dhe plotësuar një model me informacion rreth tyre. Disa fusha përdorimi janë etiketimi i tekstit, sistemet e komunikimit zanor, analizimi i të dhënave, etj. (Jurafsky & Martin, 2020; Grishman, 2014).

Gjetja e Informacionit të Kërkuar (GJIK) (angl. Information Retrieval (IR))

Qëllimi i GJIK-ë është të gjejë informacion nga një koleksion shumë i gjerë të dhënash në formë teksti. Sistemet e bibliotekës *online* dhe motorët e kërkimit në internet janë dy nga aplikimet tipike në GJIK. Në këto sisteme, si rezultat i kërkimit të informacionit të dhënat e gjetura renditen sipas rëndësisë (Mei & Radev, 2014).

Përkthimi Automatik (PA) (angl. Machine Translation (MT))

Përkthimi automatik është një nga detyrat më të rëndësishme dhe më të vjetra në përpunimin e gjuhës. Sistemet PA janë sisteme të cilat realizojnë përkthimin automatik të tekstit nga një gjuhë në një gjuhë tjetër. PA është një fushë me shumë interes si në akademi ashtu dhe në biznes. Përkthimi automatik konsiderohet si një nga detyrat më të vështira sepse këto sisteme nuk kanë për qëllim përkthimin e fjalëve nga një gjuhë në tjetrën por të përkthejnë një fjali nga një gjuhë tjetrën, duke patur si rezultat një fjali të saktë nga ana gramatikore dhe kuptimore (Specia & Wilks, 2014).

Sistemi i Dialogut në Gjuhën e Folur (SDGJF) (Spoken Language Dialogue System (SLDS))

Një sistem SDGJF është një sistem kompjuterik i aftë të realizojë një dialog zanor me një person. Këto janë sisteme të bazuara në njohjen e zërit, në leximin e tekstit ose grafikëve, apo në mjete të tjera të komunikimit. SDGJF janë sistemet e së ardhmes, që kërkojnë përdorimin e shumë teknikave të PGJN-së duke qenë se duhet të kuptojnë gjuhën dhe ta gjenerojnë atë (Dale, 2014).

Sistemet Pyetje-Përgjigje (SPP) (angl. Question-Answering (QA))

Sistemet pyetje-përgjigje janë sisteme të cilat marrin një pyetje në gjuhën natyrale nga përdoruesi dhe duhet të kthejnë një përgjigje sa më të saktë për pyetjen. Ndryshe nga sistemet GJIK ku informacioni kërkohet nëpërmjet fjalëve kyçe dhe përgjigjja e sistemit është një listë dokumentesh të indeksuar sipas rëndësisë në lidhje me fjalën kyçe, sistemet SPP marrin një pyetje, si: “Ku ndodhi ngjarja?” dhe duhet të kthejë si përgjigje një shprehje, si: “Në Tiranë.”. Pra, si rezultat kemi përsëri një shprehje në gjuhë natyrale dhe jo një apo disa dokumente (Prager, 2014).

Përmbledhja Automatike e Tekstit (PAT) (angl. Text Summarization (TS))

Sistemet e përmbledhjes automatike të tekstit duhet të krijojnë një përmbledhje me informacionet më të rëndësishme për një ose më shumë dokumente tekst të dhëna. Ka dy tipe sistemesh përmbledhjeje: përmbledhje nxjerrëse dhe përmbledhje abstrakte. Përmbledhja nxjerrëse përcakton rëndësinë e çdo fragmenti në tekstin e hyrjes dhe kthen fragmentet që kanë rëndësinë më të madhe. Përmbledhja abstrakte riformulon fragmentet e nxjerra nga teksti i hyrjes dhe i kombinon ato duke krijuar një tekst origjinal (Hovy, 2014).

KREU 3

OPINION MINING: ANALIZIMI I TEKNIKAVE QË PËRDOREN PËR KLASIFIKIMIN E OPINIONEVE

Fokusi i këtij kapitulli është analizimi fillimisht në mënyrë të përgjithshme i Opinion Mining (OM), tematikat dhe problemet që lidhen me të dhe analizimi në mënyrë të veçantë i teknikave të përdorura për detyrën e klasifikimit të opinioneve. Në këtë kapitull paraqitet një studim i thelluar i teknikave që përdoren në Opinion Mining për klasifikimin e opinioneve sipas polaritetit të ndjenjës së shprehur nga to. Këto metoda mund të klasifikohen si të mësuarit e automatizuar (MA) që ne e kemi trajtuar në çështjen 2.2 dhe të bazuara në leksik (angl. lexicon-based). Duke ju referuar publikimeve të fundit shkencore, metodat hibride, kombinim i metodave të mësuarit e automatizuar dhe të bazuara në leksik, kanë performancë më të lartë se sa kur këto metoda përdoren veç e veç. Pjesa më e madhe e punimeve shkencore në klasifikimin e opinioneve janë fokusuar në përdorimin e metodave të të mësuarit e automatizuar të kontrolluar që kërkojnë një korpus të etiketuar me madhësi të konsiderueshme, por për të eliminuar këtë kërkesë, kërkuesit gjithnjë e më shumë po fokusohen në zhvillimin e metodave të të mësuarit e automatizuar gjysmë të kontrolluar ose të pakontrolluar. Një nga problemet themelore në detyrën e klasifikimit të opinioneve sipas polaritetit të ndjenjës që ato paraqesin është që fjalët mund të paraqesin ndjenjë me polaritet të ndryshëm për fusha të ndryshme opinionesh.

Në Kote dhe Biba (2019) kemi realizuar një analizë të Opinion Mining dhe teknikave që përdoren për klasifikimin e opinioneve.

3.1. Hyrje

Opinionet dhe mendimet kanë ndikuar gjithmonë në sjelljen tonë dhe kanë një rol kyç në jetën njerëzore. Shumë nga ne para se të rezervojnë online një dhomë hoteli gjithmonë lexojnë opinionet e klientëve që e kanë vizituar atë më parë. Kështu në ditët e sotme analizimi i opinioneve për produktet, filmat, hotelet apo shërbime të tjera në mediet sociale luan një rol shumë të rëndësishëm në jetën e përditshme. Shumë biznese për të përmirësuar shërbimet e tyre janë të interesuar në aplikimin e mjeteve kompjuterike për analizimin dhe nxjerrjen e informacionit nga opinionet e klientëve të tyre.

Opinion Mining, njohur dhe si analizimi i ndjenjave (angl. Sentiment Analysis (SA)), ka si qëllim zhvillimin e metodave kompjuterike të cilat analizojnë opinionet të shprehura në formë teksti të pastrukturuar për identifikimin e mendimeve, emocioneve dhe ndjenjave që ato shprehin. Opinion Mining përfshin detyra të ndryshme ku përmendim: klasifikimi i opinioneve sipas polaritetit të ndjenjës së shprehur në to, analizimin e subjektivitetit të ndjenjave, identifikimi i aspekteve dhe entiteteve, përmbledhja e opinioneve, identifikimi i opinioneve false, etj.

Një opinion përcaktohet si “një mendim, pikëpamje, qëndrim apo gjykim i krijuar rreth një entiteti/objekti ose aspektit të entitetit/objektit nga personi që e ka shprehur atë” (Kostallari, et al., 1984). Zakonisht një opinion shprehet në mënyrë të drejtë nëpërmjet një fjalie subjektive. Por në shumë raste dhe një fjali objektive mund të shprehë në mënyrë të tërthortë një opinion nëpërmjet një fakti.

Dimensionet e klasifikimit të opinioneve mund të jenë të shumta. Duke u bazuar në mënyrën se si janë shprehur mund të klasifikohen sipas skemës së dhënë në Figurën 3.1 (Liu, 2015).

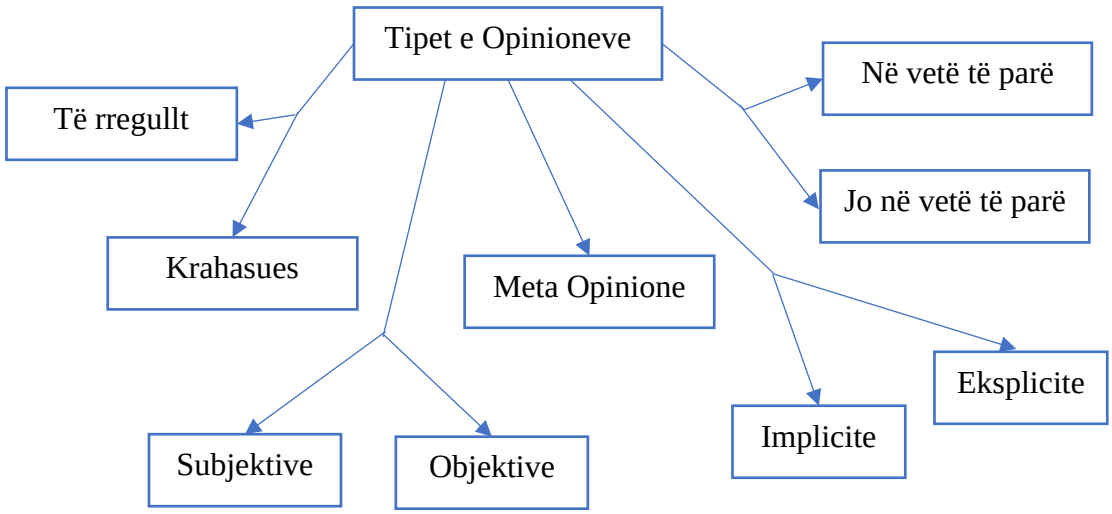


Figura 3.1 Tipet e opinioneve

Një opinion i rregullt paraqet një mendim rreth një entiteti/objekti të vetëm të shprehur direkt apo indirekt. Opinionet e shprehura në mënyrë direkte janë më thjesht të analizueshme. Kurse një opinion krahasues shpreh një mendim për një entitet/objekt duke e krahasuar me një entitet/objekt tjetër. Të dy këto tipe opinionesh mund të paraqesin një opinion subjektiv kur janë të shprehura drejtpërdrejt nëpërmjet një fjalie subjektive por mund të paraqesin dhe një opinion objektiv kur opinioni shprehet në mënyrë të tërthortë nëpërmjet një fakti. Në varësi të vetës që është shprehur opinioni mund të klasifikohet në opinione të shprehura në vetë të parë ose në opinione të shprehura në vetë të parë por për dikë në vetë të tretë. Një opinion mund të jetë një mendim rreth një opinioni tjetër dhe këto quhen meta opinione. Në këto lloj opinionesh objektivi i opinionit është vet një opinion i shprehur nëpërmjet një fjalie të përbërë varësie.

Një opinion mund të përcaktohet si një set të përbërë nga pesë elementë, (*e*, *a*, *n*, *p*, *k*), ku *e* është entiteti për të cilin është shprehur opinioni, *a* është një aspekt i entitetit *e* për të cilin shprehet opinioni, *n* është polariteti i mendimit të opinionit të shprehur për aspektit *a* të entitetit *e*, *p* është personi që shpreh opinionin, dhe *k* është koha kur opinioni shprehet nga *p*, ku p.sh. n mund të jetë pozitive, negative, neutrale ose një vlerësim me yje, 1-5 yje,

etj. (Liu, 2015). Bazuar në këtë përkufizim, detyra e klasifikimit të opinioneve sipas polaritetit të ndjenjës së shprehur nga ai në OM duhet të vlerësojë dhe përcaktojë pesë elementët e opinionit.

Opinionet mund të klasifikohen në nivele të ndryshme duke u bazuar në polaritetin e ndjenjës së shprehur. Mund të kemi klasifikim në dy klasa: pozitive ose negative, ku një opinion klasifikohet si pozitiv nëse polariteti i ndjenjës së shprehur është pozitiv dhe si negativ nëse polariteti i ndjenjës së shprehur është negativ. Por mund të kemi dhe klasifikime me më shumë klasa: si pozitive, negative dhe neutrale; ose klasifikime me yje apo pikë.

Detyra e klasifikimit të opinioneve sipas polaritetit të ndjenjës së shprehur mund të konsiderohet si detyra tradicionale e klasifikimit të dokumenteve sipas temave. Në klasifikimin tradicional të dokumenteve tekst sipas temave që trajton dokumenti janë shumë të rëndësishme fjalët specifike të temës, kurse në rastin e klasifikimit të opinioneve mbi bazën e polaritetit të ndjenjës së shprehur, fjalët që shprehin nëse opinioni është pozitiv ose negativ kanë më shumë rëndësi, të tilla si: shumë i bukur, i tmerrshëm, jofunksional, i mirë, etj.

Teknikat më të përdorshme për klasifikimin e opinioneve janë teknikat e të mësuarit të kontrolluar, të cilat mësojnë njohuri nga korpuse të dhënash të etiketuara. Në varësi të korpuseve të opinioneve, opinion mining mund të jetë: *in-domain opinion mining* kur korpusi i opinioneve të trajnimit dhe të testimit janë të të njëjtës fushë; *multi-domain opinion mining* kur të dy korpuset përbëhen nga opinione të fushave të ndryshme, *cross-domain opinion mining* kur korpusi i trajnimit është nga një fushë dhe korpusi i testimit është nga një fushë tjetër dhe *cross-lingual opinion mining*, kur korpuset trajnimit dhe testimit të opinioneve janë në gjuhë të ndryshme.

Shumica e punimeve shkencore sugjerojnë që opinionet tekst përpara se të përdoren për të trajnuar dhe testuar një model për qëllime të ndryshme të kalojnë në një fazë parapërpunimi që ndikon në përmirësimin e modelit. Disa nga këto metoda që mund të përmendim janë n-gram, TF-IDF, përdorimi i etiketave të pjesëve të ligjëratës (angl. part-of-speech tag), gjetja e temës (angl. lemmatization) ose gjetja e rrënjës (angl. stemming), etj.

3.2. Detyrat dhe aplikimet e Opinion Mining

Opinion Mining përfshin një gamë të gjerë aplikimesh dhe detyrash. Në Figurën 3.2 tregohen detyrat kryesore që realizohen në OM dhe në Figurën 3.3 tregohen disa nga fushat e aplikimit të OM.

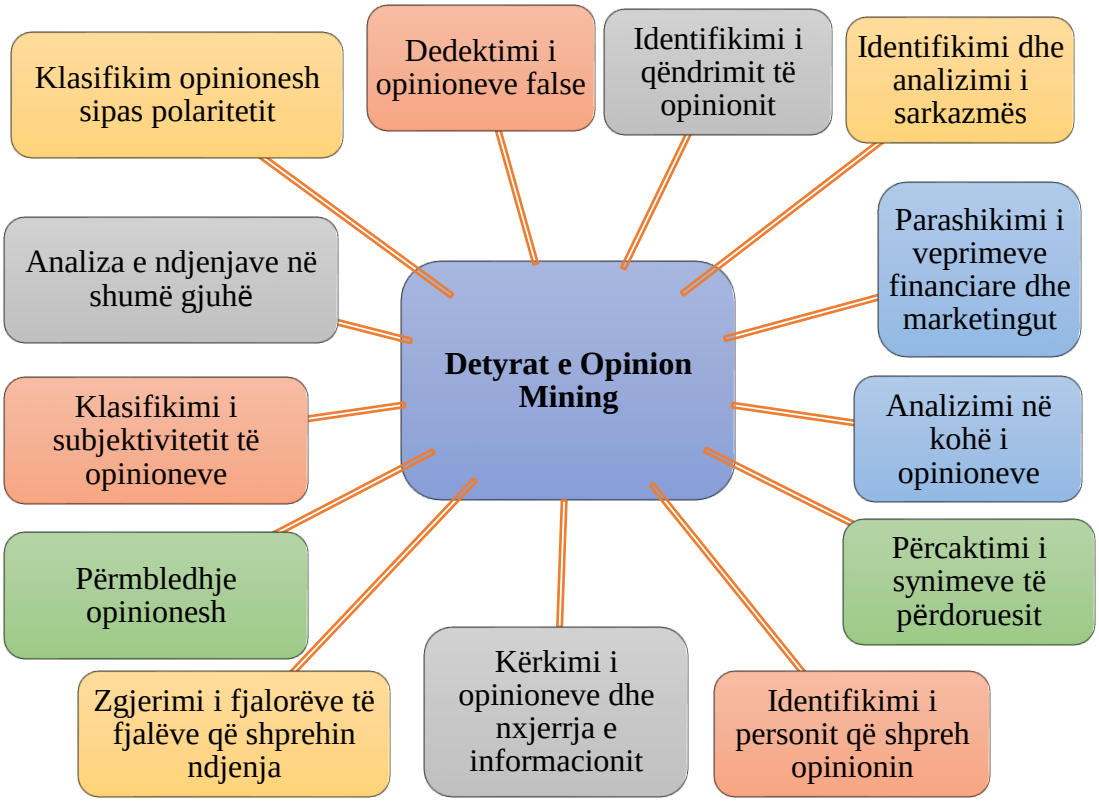


Figura 3.2 Detyrat e Opinion Mining

Detyra e klasifikimit të opinioneve ka si qëllim të klasifikojë opinionet sipas një grupi kategorish të caktuara duke u bazuar në polaritetin e ndjenjës të shprehur në to. Në këtë kapitull është trajtuar në mënyrë të detajuar kjo detyrë prandaj nuk po ndalemi këtu në dhënien e shembujve të aplikimit të kësaj detyre.

Nxjerrja e informacionit është një tjetër detyrë e OM që ka si qëllim të identifikojë dhe të nxjerrë informacionet e nevojshme nga një opinion. Një nga aplikimet më të përhapura është identifikimi i aspekteve të ndryshme të një produkti apo shërbimi të cilat theksohen në opinion dhe informacioni që lidhet me to. Ky është niveli më i thellë i analizimit të opinioneve. Një metodë e thjeshtë për nxjerrjen e aspekteve të produkteve dhe mendimit të shprehur rreth tyre është përdorimi i të mësuarit të thelluar (angl. deep learning) nëpërmjet një rrjeti të thjeshtë CNN që përshin dy tipe embeddings të para-trajnuara: embeddings me qëllim të përgjithshëm dhe embeddings specifikë nga tema e opinionit. Kjo metodë e thjeshtë ka rezultate të mira dhe performon më mirë se metoda ekzistuese më të sofistikuara (Xu, et al., 2018).

Një tjetër detyrë është analizimi i subjektivitetit të opinioneve, që ka si qëllim të identifikojë nëse një fjali në një opinion shpreh një opinion ose jo. Ky analizim i opinionit realizohet në nivel fjalie (Liu, 2010).

Identifikimi i opinioneve false ka si qëllim të identifikojë nëse një opinion është fals ose jo. Nëpërmjet përdorimit të sistemeve të cilat identifikojnë nëse një opinion i shprehur në mediet sociale është apo jo fals dhe duke analizuar mendimet e shprehur në këto opinione rreth produkteve të ndryshme, kompanitë dhe menaxherët mund të nxjerrin informacione shumë të nevojshme për të marrë vendimet e duhura në të ardhmen për rritjen e biznesit të tyre (Kauffmanna, et al., 2020).

Ndërtimi i aplikacioneve të cilat gjenerojnë përmbledhje të opinioneve në mediet sociale, sidomos në një fushë të veçantë si në hoteleri, ka tërhequr shumë vëmendje vitet e fundit. Në modele të tilla ka një rendësi të veçantë identifikimi dhe klasifikimi i opinioneve sipas rëndësisë, identifikimi i aspekteve që lidhen me temën e opinionit dhe klasifikimi i polaritetit të ndjenjës së shprehur. Mbi bazën e këtyre elementëve mund të krijohen përmbledhje të sakta të opinioneve për një temë të caktuar (Tsai, et al., 2020).

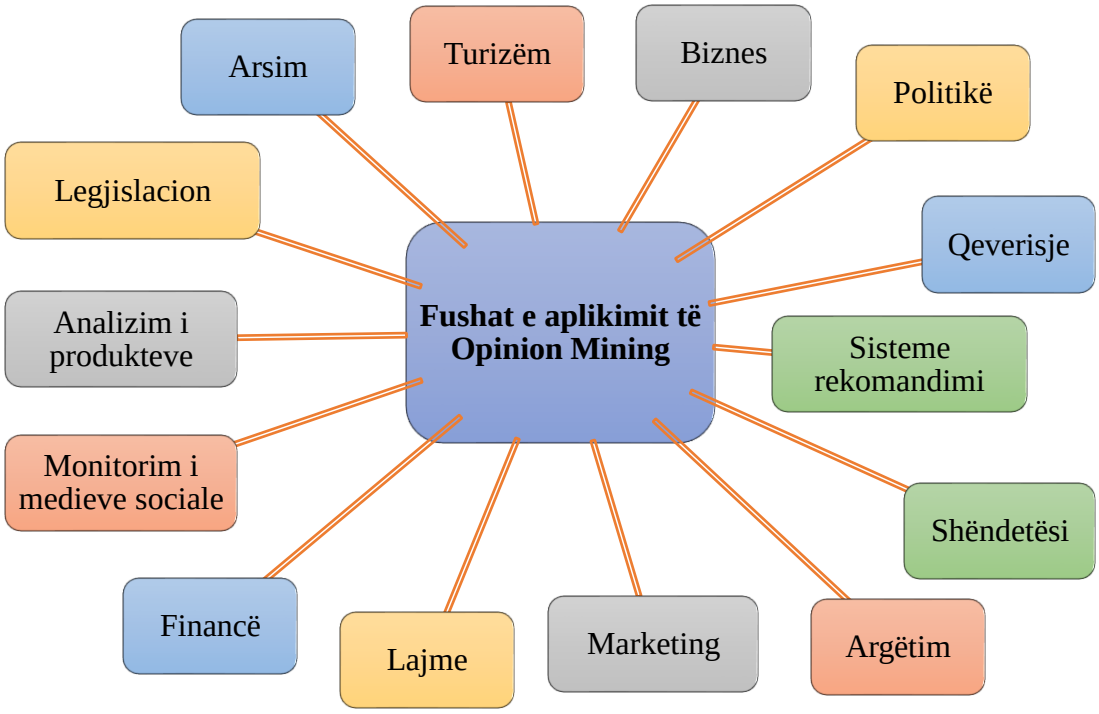


Figura 3.3 Fusha të aplikimit të Opinion Mining

Analizimi i opinioneve dhe i ndjenjave ka aplikim të gjerë në fusha ku duhet të kuptojmë mendimet e qenieve njerëzore. Një nga aplikimet tipike dhe më të përhapura sot është në fushën e biznesit. Bizneset përdorin këto teknika për të analizuar opinionet e klientëve për produktet apo shërbimet e tyre të cilat i ndihmojnë për të marrë vendime strategjike në të ardhmen.

Një tjetër aplikim është përdorimi i këtyre metodave nga qeveria për të identifikuar dhe kuptuar atë që qytetarët vërtet kanë nevojë dhe duan. Në punimin e tyre Arunachalam dhe Sarkar (2013) kanë propozuar një model për analizimin e mendimeve dhe ndjenjave të qytetarëve në rrjetet sociale nga ana e qeverisë. Modeli i propozuar është implementuar në probleme reale të një agjencie qeveritare Amerikane duke nxjerrë informacione të vlefshme që mund të shërbejnë në përcaktimin e strategjive që do të ndjekë qeveria në të ardhmen.

Aplikim të gjerë OM ka dhe në fushën politike, sidomos në parashikimin e rezultateve të zgjedhjeve mbi bazën e analizimit të polaritetit të opinioneve të shprehura nga elektorati online. Tsakalidis et al. (2015) kanë përdorur teknikat OM për të analizuar dhe parashikuar rezultatet e zgjedhjeve parlamentare të Komisionit Evropian të vitit 2014 nëpërmjet analizimit të opinioneve të shprehur në Twiter.

Analizimi i ndjenjave të shprehura në opinionet në mediet sociale dhe jo vetëm, në fushën e turizmit ka patur një vëmendje gjithmonë e në rritje vitet e fundit. Mjaft punë kërkimore kanë propozuar aplikacione të cilat analizojnë ndjenjat e shprehura në opinione për aspekte të ndryshme të turizmit. Kirilenko et al. (2018) kanë analizuar dhe krahasuar në mënyrë të detajuar metoda të ndryshme të përdorura për analizimin e opinioneve në fushën e turizmit duke i krahasuar ato me aftësinë humane për analizim të opinioneve. Metoda e tyre krahasuese u vjen në ndihmë shumë kërkuesve të cilët kanë si qëllim të përmirësojnë teknikat ekzistuese apo dhe kompanive të ndryshme që duan të analizojnë opinionet e klientëve të tyre duke përdorur teknikën më eficiente. Në përfundimet e tyre autorët theksojnë se këto metoda automatike të klasifikimit të ndjenjave kanë performancë të krahasueshme me aftësitë njerëzore në klasifikimin e ndjenjave. Një faktor që ndikon negativisht në performancën e këtyre metodave janë të dhënat komplekse dhe ato që përmbajnë informacione të panevojshme që konsiderohen si të dhëna që përmbajnë zhurmë. Aplikimet më të përdorshme të analizës së opinioneve dhe ndjenjave në fushën e turizmit janë: monitorimi i biznesit, produkteve apo shërbimeve me kalimin e kohës, strategjitë inteligjente dhe konkurruese, trendet në nivel makro, menaxhimi i rrezikut të reputacionit dhe përgjigjja e pyetjeve të ndryshme të me anë të analizimit të të dhënave për një vendimmarrje sa më të mirë për zhvillimin e biznesit në të ardhmen, etj. (Thelwall, 2019).

3.3. Nivelet e analizimit të opinionit

Qëllimi i OM është të zhvillojë mjete që të nxjerrin dhe të analizojnë të dhëna subjektive nga opinione tekst të shkruar në gjuhën natyrale. Në përgjithësi analizimi dhe klasifikimi i opinioneve duke u bazuar në polaritetin e ndjenjës së shprehur në to mund të realizohet në tri nivele: në nivel dokumenti, në nivel fjalie dhe në nivel aspekti dhe entiteti.

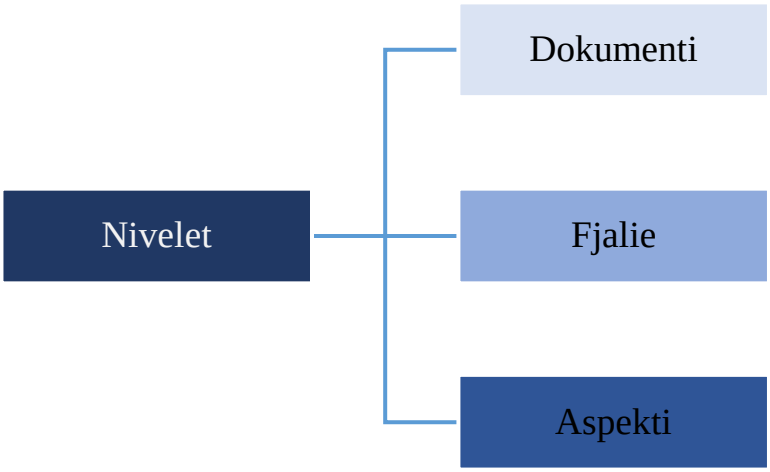


Figura.3.4 Nivelet e analizimit të opinioneve

3.3.1. Nivel dokumenti

Analizimi në nivel dokumenti është niveli më i lartë i abstraktimit të analizimit të opinioneve. Në këtë nivel, opinioni konsiderohet si një entitet i vetëm dhe qëllimi është të klasifikohet duke përcaktuar polaritetin e ndjenjës së shprehur nga ai në tërësi, pa analizuar në detaje aspektet, entitetet apo mendimet e veçanta të shprehura në të. Klasifikimi i opinioneve në këtë nivel mund të konsiderohet si klasifikim tradicional i dokumenteve tekst.

Në klasifikimin e opinioneve në nivel dokumenti supozohet që një opinion i ruajtur në një dokument tekst shpreh një opinion për një entitet/objekt të vetëm dhe ky opinion është shprehur nga një person *p* i vetëm. Detyra e klasifikimit në këtë rast është të përcaktohet polariteti i ndjenjës së shprehur nga opinioni për entitetin *e* duke përcaktuar polaritetin e mendimit *n* për aspektin *a* në (*e*, *a*, *n*, *p*, *k*), ku *a*=*e*, pra entiteti në këtë rast është edhe aspekti për të cilin shprehet mendimi dhe *n*, *p* dhe *k* janë të njohur ose të panjohura (Liu, 2015). Ky nivel klasifikimi nuk është i përshtatshëm për opinionet krahasuese ose opinionet që shprehin mendime rreth disa objekteve duke qenë se aspekti për të cilin përcaktohet polariteti i mendimit është një entitet i vetëm. Në rastin e opinioneve krahasuese, opinioni shpreh mendim për një entitet duke e krahasuar me një entitet tjetër dhe nuk mund të përcaktojmë nëse mendimi është pozitiv apo negativ.

3.3.2. Nivel fjalie

Analizimi i opinioneve në nivel dokumenti është shumë i përgjithshëm, dhe për këtë arsye kërkuesit janë përqendruar në metoda për një analizë më të detajuar në nivel fjalie.

Liu (2015) përcakton që në klasifikimin e opinioneve në nivel fjalie, për një fjali të dhënë duhet të realizohen dy detyra:

1. Të përcaktohet nëse fjalia është një fjali subjektive apo objektive;

2. Nëse fjalia është një fjali subjektive, të përcaktohet polariteti i ngjyrimit të mendimit të shprehur nga fjalia, si pozitive ose negative.

Të dyja detyrat e mësipërme konsiderohen si detyra klasifikimi. Supozimi i bërë në këtë rast është që fjalia subjektive të shprehë një opinion për një entitet të caktuar. Ky supozim nuk vlen në rastin e fjalive të përbëra ose fjalive që shprehin opinione rreth aspekteve të ndryshme të një entiteti apo fjalive që përmbajnë një opinion krahasues. Në këto raste në varësi të polaritetit të mendimit të shprehur për entitete apo aspekte të ndryshme mund të përcaktohet nëse fjalia opinion shpreh një polaritet të caktuar, pozitiv apo negativ.

Nëse polariteti i ngjyrimit të mendimit të një opinioni të shprehur me më shumë se një fjali do të analizohej në nivel fjalie atëherë për çdo fjali të opinionit fillimisht duhet të përcaktohet nëse ajo është subjektive apo objektive. Më pas vetëm për fjalitë subjektive përcaktohet polariteti i mendimit që shpreh fjalia. Polariteti i mendimit të opinionit në tërësi (të përbërë nga shumë fjali), në nivel dokumenti, vlerësohet si bashkim i polariteteve të çdo fjalie (Pang & Lee, 2008).

Liu (2015) duke u bazuar në faktin që edhe fjalitë objektive mund të shprehin një opinion për një objekt nëpërmjet një fakti, sugjeron që si fjalitë objektive ashtu edhe ato subjektive duhet të merren parasysh në përcaktimin e polaritetit të një opinioni si në nivel fjalie ashtu dhe në nivel dokumenti.

3.3.3. Nivel aspekti dhe entiteti

Në klasifikimin në nivel aspekti dhe entiteti, opinioni analizohet në mënyrë më të detajuar dhe identifikohen aspektet e entitetit apo entitetet e ndryshme për të cilët shprehet opinioni dhe më pas analizohet polariteti i mendimit të shprehur për secilin entitet apo aspekt të entitetit të identifikuar. Në këtë mënyrë një opinion mund të klasifikohet si pozitiv për një aspekt të një entiteti apo një entitet por negativ për një aspekt të një entiteti apo të një entiteti tjetër. Në këtë mënyrë do të shmanget përcaktimi i një polariteti të përgjithshëm si në rastin e klasifikimit në nivel dokumenti ku jo domosdoshmërisht polariteti i përcaktuar mund të përkojë me polaritetin e mendimit për çdo aspekt apo entitet që është shprehur një mendim.

Nëse marrim në konsideratë përcaktimin e opinionit si në set nga pesë elementët (*e*, *a*, *n*, *p*, *k*), klasifikimi i opinionit në nivel aspekti dhe entiteti fokusohet në realizimin e dy detyrave:

1. Identifikimi i entitetit *e* dhe/ose aspektit *a* të entitetit *e* për të cilën shprehet një mendimi në opinion;
2. Përcaktimi nëse opinionin për aspektin *a* të një entiteti *e* ose për një entitet të caktuar *e* të identifikuar shpreh një ndjenjë *n*, negative apo pozitive.

Klasifikimi në këtë nivel analizon në mënyrë të detajuar opinionin duke konsideruar opinionin në vetvete në vend të strukturës gjuhësore të tij (Liu, 2015).

3.4. Teknikat e klasifikimit të opinioneve

Në këtë pjesë janë analizuar teknikat që përdoren për detyrën e klasifikimit të opinioneve sipas polaritetit të tyre. Teknikat e përdorshme për këtë qëllim klasifikohen në teknika të të mësuarit të automatizuar (MA) (angl. Machine Learning, (ML)) dhe të bazuara në leksik (angl. lexicon-based).

3.4.1. Teknikat e të mësuarit e automatizuar

Siç kemi diskutuar në çështjen 2.2, të mësuarit e automatizuar (MA) është një fushë e Inteligjencës Artificiale (IA) që studion teknika dhe algoritme që mësojnë njohuri nga të dhënat dhe parashikojnë rezultate për të dhëna të reja. Në rastin e OM, një model realizohet nga një algoritëm i cili mëson polaritetin e opinioneve duke u trajnuar nëpërmjet një korpus opinionesh të etiketuar, si p.sh. opinione pozitive dhe negative ose identifikon polaritetin e opinionit nga një korpus i paetiketuar dhe njohuritë e fituara përdoren për parashikimin e polaritetit të opinioneve të reja. Duke u bazuar në mënyrën e të mësuarit që kemi diskutuar në çështjen 2.2 teknikat e të mësuarit e automatizuar i kemi klasifikuar në teknika të të mësuarit të kontrolluar (angl. supervised learning) dhe teknika të të mësuarit të pakontrolluar (angl. unsupervised learning). Teknikat e të mësuarit gjysmë të kontrolluar (angl. semi-supervised learning) janë teknika të cilat përdorin teknika si të të mësuarit të kontrolluar dhe ato të pakontrolluar. Këto teknika kanë përdorim të gjerë në Opinion Mining dhe prandaj i kemi trajtuar.

3.4.1.1. Teknikat e të mësuarit të kontrolluar

Teknikat e të mësuarit të kontrolluar përdoren për të ndërtuar modele duke mësuar njohuri nga të dhëna të etiketuara për atribute/karakteristika të caktuara dhe i përdorin ato për parashikimin e atributeve/karakteristikave për të dhëna të reja të paetiketuara. Në Opinion Mining, një model krijohet duke trajnuar një algoritëm nëpërmjet një korpusi opinionesh të etiketuar për polaritetin të mendimit të opinioneve si p.sh pozitiv dhe negativ. Modeli i trajnuar përdoret për parashikimin e polaritetit të ndjenjës të opinioneve të reja. Performanca e algoritmeve MA varet nga sasia dhe cilësia e të dhënave të etiketuara me të cilat trajnohen. Rritja e sasisë së të dhënave trajnuese rrit performancën, po në të njëjtën kohë rrit dhe koston e përdorimit të tyre (Choi & Lee, 2017).

Pang et al. (2002) ishin të parët të cilët propozuan përdorimin e teknikave MA, që deri tashme ishin përdorur për klasifikimin e tekstit sipas fushave, për klasifikimin e polaritetit të mendimit të opinioneve. Ata propozuan përdorimin e të njëjtës teknikë si në klasifikimin tradicional dhe në klasifikimin e opinioneve të filmave për t'i klasifikuar sipas polaritetit të ndjenjës së shprehur nga opinioni. Në eksperimentet e zhvilluara ata vlerësuan performancën e tri algoritmeve të të mësuarit të kontrolluar Naive Bayes, SVM dhe Maximun Entropy në një korpus opinionesh filmash të klasifikuar në dy nivele si pozitive dhe negative. Rezultatet tregojnë që këto algoritme në klasifikimin e opinioneve kanë

performancë më të ulët se sa në klasifikimin tradicional të tekstit dhe ka shumë sfida që duhen adresuar në të ardhmen. Algoritmi SMV ka performancën më të mirë dhe Naive Bayes më të ultën. Përdorimi i unigram është metoda që ka rezultuar më efektive.

Naive Bayes (NB)

NB është një nga klasifikuesët më të thjeshtë dhe popullor në klasifikimin e teksteve. NB ka eficiencë të lartë dhe kërkon një bashkësi të vogël të dhënash për t'u trajnuar. Ky algoritëm i konsideron karakteristikat e përdorura të pavarura nga njëra-tjetra dhe nëse në detyrën e klasifikimit njëra nga klasat ka më shumë të dhëna se sa tjetra, algoritmi zgjedh pesha të dobëta për vijën ndarëse. Këto probleme anashkalohen në algoritmin e përmirësuar Multinomial Naive Bayes, i cili e modelon shpërndarjen e fjalëve në dokument si një multinomial. Në këtë rast teksti konsiderohet si një *bag-of-words* dhe pozicioni i çdo fjale është i pavarur nga fjalët e tjera. Për të realizuar klasifikimin e të dhënave tekst supozohet që ka një numër fiks klasash, secila me një numër fiks parametrash multinomial. Algoritmi Multinomial Naive Bayes është shumë i shpejtë, i thjeshtë për t'u implementuar dhe ka performancë të mirë në klasifikimin e tekstit (Rennie, et al., 2003).

Në punimin e tyre, Dinu dhe Iuga (2012), janë fokusuar në dy çështje kryesore për sa i përket përdorimit të algoritmit Naive Bayes për klasifikimin e opinioneve të filmave në dy klasa: cilat janë karakteristikat që duhen patur parasysh gjatë trajnimit të një modeli duke përdorur këtë algoritëm dhe nëse kombinimi i modeleve të trajnuara me karakteristika të ndryshme sjell rezultate më të mira në parashikim. Karakteristikat e vlerësuara në këtë rast janë: fshirja e parafjalëve, gjetja e rrënjës së fjalëve, bigram dhe gjithë fjalët, përdorimi në trajnim i fjalëve më të përdorshme duke i kombinuar ose jo me bigram, përdorimi i etiketave të pjesëve të ligjëratës, shtimi i sinonimeve në korpus. Rezultatet eksperimentale tregojnë që kombinimi i disa karakteristikave si bigram, përdorimi i etiketave të pjesëve të ligjëratës dhe shtimi i sinonimeve rezultuan në performancë më të mirë të algoritmit në parashikimin e polaritetit të ndjenjës së opinionit.

Algoritmi NB është përdorur për të realizuar një sistem i cili analizon opinionet e klientëve për produkte në nivel fjalie dhe aspekti në dy klasa, pozitive dhe negative. Vlerësimet eksperimentale tregojnë që modeli i propozuar ka performancë të mirë në identifikimin e aspekteve dhe polaritetit të tyre (Jeyapriya & Selvi, 2015).

Support Vector Machines (SVM)

SVM është një algoritëm linear klasifikimi që kombinon në mënyrë lineare karakteristikat e të dhënave për të përcaktuar klasën që ato i përkasin. Ky algoritëm ka përdorim shumë të gjerë në klasifikimin e opinioneve sipas polaritetit të ndjenjës së shprehur. Një model SMV i përfaqëson të dhënat si pika në hapësira në mënyrë të tillë që të dhënat që i përkasin kategorive të ndryshme të jenë të ndara nga njëra-tjetra me një hapësirë sa më të madhe të mundshme. Të dhënat e reja parashikohen se i përkasin një kategorie të caktuar duke u bazuar në cilën anë të hapësirës ato kanë rënë.

Në Weka për të ndërtuar një model klasifikimi mbi bazën e SVM përdoret algoritmi Sequential Minimal Optimization (SMO). Algoritmi SMO përdor një procedurë përsëritëse të katrorëve më të vegjël të ripeshuar për optimizim dhe një strategji të rastësishme përzgjedhëse midis të dhënave të përdorura gjatë trajnimit.

Për të zgjidhur detyra të mëdha kuadratike në programim, SMO e ndan atë në sekuenca më të vogla detyrash kuadratike. Këto detyra të vogla zgjidhen në mënyrë analitike më tej për të shmangur përdorimin e teknikave të optimizimit që kërkon shumë kohë. Algoritmi SMO kërkon që gjatë fazës së trajnimit të përdoren sasi shumë të mëdha të dhënash duke qenë se memoria e të mësuarit është lineare me madhësinë e të dhënave. Kjo është dhe një nga problemet e përdorimit të këtij algoritmi, sepse krijimi i korpuseve të mëdha të etiketuara të trajnimit kërkon aftësi të mira njerëzore në etiketim dhe kohë. Rezultatet e një vlerësimi krahasues tregon se algoritmi SMO mund të jetë më i shpejtë se algoritmi i copëzimit PCG për SVM lineare (Platt, 1999).

Algoritmi SMV është përdorur nga Severyn et al. (2015) për të trajnuar një model për parashikimin e polaritetit të mendimit dhe tipit të opinionit për opinione të marra nga YouTube rreth dy produkteve: automjete dhe tableta. Sistemi i prezantuar në këtë punim është vazhdim i dy punime të mëparshëm: Uryupina et al. (2014) në të cilin prezantohet korpusi i etiketuar dhe Severyn et al. (2014) në të cilin prezantohet sistemi në mënyrë empirike. Autorët adresojnë problemin që opinionet në Youtube mund të shprehin mendim rreth video dhe/ose produktit, për këtë arsye ata propozojnë një sistem me shumë klasa klasifikimi. Modeli i propozuar përdor skemën "një-kundrejt-të-gjithave" për të përcaktuar polaritetin e ndjenjës së opinionit, nëse komenti është për videon apo produktin. Gjithashtu është realizuar dhe një eksperiment cross-domain për të adaptuar njohuritë që modeli mëson duke u trajnuar me opinione rreth një produkti të caktuar për të klasifikuar opinione për një produkt tjetër. Më tej propozimi është përdorur për të realizuar një model duke përdorur një korpus trajnimi opinionesh të po këtyre produkteve në gjuhën italiane. Qëllimi i këtij eksperimenti të fundit ka qenë të vlerësohet performanca e propozimit për gjuhë që kanë burime të pakta në fushën e përpunimit të gjuhës natyrale. Vlerësimet eksperimentale tregojnë që modeli i propozuar performon shumë mirë dhe më mirë se modeli *bag-of-words*.

Algoritmi SVM raportohet të ketë performancë të mirë në klasifikimin e polaritetit të ndjenjës së opinioneve për produkte të ndryshme dhe nga modeli i prezantuar në punimin e Saleh et al. (2011). Gjatë eksperimenteve janë përdorur karakteristika të ndryshme si frekuenca e termit - frekuenca e anasjelltë e dokumentit (angl. Term Frequency — Inverse Document Frequency (FT-IDF)), përhapja binare (angl. Binary Occurrence (BO)), përhapja e termit (Term Occurrence (TO)) dhe n-gram për të identifikuar sa ndikojnë këto karakteristika në perforfmancën e modelit. Rezultatet më të mira merren kur përdoren karakteristikat TFIDF me bigram. Përdorimi i procedurave të peshuara nuk rrit performancën e modelit madje ka rezultate më të ulëta. Rezultat e arritura tregojnë se kjo metodë ka performancë të lartë.

Të dy këto algoritme NB dhe SVM janë tepër të suksesshëm në klasifikimin e tekstit dhe opinioneve. Performanca e modeleve të të dy algoritmeve në klasifikimin e opinioneve varet nga karakteristikat e përdorura, në rastin e përdorimit të bigram perdormanca është më e mirë dhe nga korpusi. Algoritmi NB ka performancë më të mire në Opinion Mining për opinione tekst të shkurtra kurse SVM për opinione tekst të gjata. Modeli i bazuar në algoritmin SVM që përdor raportet log-count të algoritmit NB ka një performancë më të lartë se rastet e modeleve të veçanta të algoritmeve (Wang & Manning, 2012).

Bayesian Network (BN)

Rrjeti BN është një model grafik probabilitar. Nga ana strukturore, BN është një graf aciklik i orientuar, ku nyjet përfaqësojnë variablat dhe harqet përfaqësojnë lidhjen midis nyjeve, pra variablave. Nëse kemi një hark nga një nyje A në një nyje B do të thotë që nyja A është prindi i nyjës B, ku nyja A mund të jetë një varibël i çfarëdoshëm. Një BN me parametra është një paraqitje grafike e shpërndarjes së përbashkët mbi të gjitha variablat e përfaqësuar nga nyjet në grafik. Nëse variablat janë X1, ..., Xn dhe "prindër (A)" të jenë prindër i nyjës A. Shpërndarja e përbashkët për variablat X1 deri në Xn paraqitet si një produkt i shpërndarjes probabilitare:

$$P(X1, \dots , Xn) = P(Xi \text{ prindi } (Xi)) \text{ për } i = 1 \text{ në } n.$$

Në këtë mënyrë për të specifikuar një BN duhet të specifikohet për çdo nyje shpërndarja probabilitare për X kushtëzuar nga prindi i tij. Në këtë mënyrë ky rrjet do të përdoret për të llogaritur vlerën e probabilitetit mbi variablat e të dhënave (Russell & Norvig, 2020).

Rrjetet BN janë përdorur gjerësisht në krijimin e modeleve për klasifikimin e polaritetit të mendimeve të opinioneve por gjithashtu janë përdorur edhe si mjete ndihmëse për nxjerrjen e karakteristikave ndërmjet variablave të të dhënave për modele të bazuar në algoritme të tjerë (Gutiérrez, et al., 2018).

Një model i bazuar në një rrjet BN është propozuar nga Ren dhe Kang (2013) për të identifikuar jo vetëm emocionet e thjeshta por dhe emocione më komplekse në opinione tekst. Modeli bashkon analizën e emocioneve me modelin e fushës për të identifikuar emocione komplekse, intensitetin e emocionit dhe ndryshimin e emocionit në varësi të fushës në dokumente tekst. Rrjeti BN përdoret për gjetjen e variablave të fshehur të fushës dhe të emocioneve. Rezultatet eksperimentale tregojnë se ky model është më efektiv se klasifikuesit tradicionalë Naive Bayes dhe SVM në rastin e emocioneve të thjeshta dhe ka rezultate premtuese dhe në rastet komplekse të analizimit të emocioneve. Gjithashtu autorët theksojnë se rezultatet tregojnë që ka një lidhje midis temës së dokumenteve dhe emocioneve që shprehen nga to.

Maximum Entropy (ME)

ME është një algoritëm i cili njihet dhe si një model eksponencial duke qenë se ka një formulë eksponenciale. Në ndryshim nga Naive Bayes, ME nuk supozon që karakteristikat janë të pavarura nga njëra-tjetra. Habernal et al. (2014) kanë realizuar një analizë të detajuar të performancës së modeleve të bazuar në ME duke e krahasuar me modele të bazuara në SVM për klasifikimin e opinioneve në gjuhën çeke të klasifikuara në tri klasa si pozitive, negative dhe neutrale duke përdorur karakteristika dhe teknika të ndryshme të parapërpunimit të tekstit, si n-gram, etiketat e pjesëve të ligjëratës, gjenerimi i temës së fjalëve, etj. Rezultatet eksperimentale tregojnë se ME në shumicën e rasteve ka performancë më të mirë se SVM.

Algoritmi *SGD* përdoret për të implementuar një zbritje të gradientit stokastik për të krijuar një model linear të të mësuarit si një regresion logjistik me klasë binare dhe klasa binare SVM. Atributet nominale konvertohen në atribute binare. Atributet janë të normalizuara dhe rrjedhimisht dhe të dhënat janë të normalizuara. Dalja varet nga këto të dhëna të normalizuara (Witten, et al., 2017).

Në punimin e Tripathy et al. (2016) është vlerësuar performanca e katër algoritmeve NB, ME, SGD dhe SVM duke përdorur teknikat e n-gram-eve, TF-IDF dhe CountVectorizer për klasifikimin e opinioneve në dy klasa pozitive dhe negative. Rezultatet eksperimentale tregojnë që saktësia e klasifikimit është më e mirë kur përdoret 1-gram dhe 2-gram dhe përdorimi i teknikave TF-IDF dhe Count Vectorizer përmirëson saktësinë dhe modeli i ndërtuar duke përdorur algoritmin SVM ka perfomancën më të mirë në këtë rast.

Logistic Regression (LR)

Algoritmet LR janë algoritme statistikore që përdoren për të modeluar të dhëna binare duke përdorur një funksion logjistik. Për secilën klasë, funksioni logjistik llogarit vlerën e probabilitetit dhe më pas zgjidhet klasa që i korrespondon vlerës maksimale të probabilitetit (Landwehr, et al., 2005). Algoritmet LR të marra në konsideratë për klasifikimin e opinioneve në gjuhën shqipe në këtë studim janë Logistic dhe Simple Logistics të implementuar në Weka. Algoritmi Logistic i implementuar në Weka është një modifikim i algoritmit origjinal LR për të trajtuar peshat e instancave. Për të rritur performancën e algoritmit dhe për të ulur gabimin në parashikim, ky algoritëm mund të përdoret bashkë me vlerësuesit e kreshtës (angl. ridge estimators) (Cessie & Houwelingen, 1992). Algoritmi Simple Logistic përdor algoritmin Logit Boost për të ndërtuar një model multinomial LR. Ky algoritëm përdoret të përshtatur modelin dhe është përcaktuar një numër optimal iteracionesh në algoritëm për të përzgjedhur automatikisht atributet. Ky numër optimal iteracionesh është i rëndësishëm sepse në të kundërt nëse do të ekzekutohej algoritmi me një numër shumë të madh iteracionesh do të merrej i njëjti model sikur të ishte përdorur algoritmi Logistic (Sumner, et al., 2005).

Bayesian Logistic Regression (BLR)

Algoritmi BLR është implementimi i një metode Bayesiane për regresion logjistik. Ky algoritëm përdor shpërndarjen Laplace për të shmangur mbingopjen e modelit dhe për të ndërtuar një model të shpërndarë parashikues për të dhëna tekst. Aplikimi i këtij algoritmi për klasifikimin e dokumenteve tregon se performanca e tij është e krahasueshme me performancën e algoritmeve SVM dhe LR, prandaj ky algoritëm mund të ketë rezultate të mira dhe për klasifikimin e opinioneve sipas polaritetit të mendimit të shprehur (Genkin, et al., 2007).

K-nearest neighbor (K-NN)

është një algoritëm i thjeshtë që përdoret për dy qëllime: për klasifikimin e të dhënave dhe për të vlerësuar funksionin e shpërndarjes së densitetit të të dhënave. Algoritmi supozon që të dhënat janë pika në një hapësire n-dimensionale. K është një numër i vogël i rastësishëm i fqinjëve të një pike dhe distanca e fqinjësisë përcaktohet mbi bazën e intervaleve Euklidiane. Kur ky algoritëm përdoret për klasifikim, e dhëna e re krahasohet me çdo të dhënë në korpusin e trajnimit dhe klasa e saj përcaktohet duke u bazuar në distancën nga k-fqinjëve më të afërt të tij. Klasat e fqinjëve peshohen bazuar në ngjashmërinë e të dhënave të reja me secilin fqinj dhe distanca Euklidiane përdoret për të matur ngjashmërinë. Në artikullin e İşgüder-Şahin et al. (İşgüder-Şahin, et al., 2014), algoritmi K-NN është përdorur për të ndërtuar një model për klasifikimin e opinioneve në gjuhën turke në dy klasa, pozitive dhe negative. Algoritmi IBk është implementimi i një algoritmi K-NN në Weka (Witten, et al., 2017). Në ndërtimin e modelit është përdorur karakteristika e bigram-eve e cila ndikon në përmirësimin e performancës së modelit. Modeli i ndërtuar me algoritmin K-NN është krahasuar për të njëjtën detyrë me dy model të tjera të bazuar në algoritmin NB dhe SVM. Rezultatet eksperimentale tregojnë që modeli i SVM ka performancë më të mirë krahasuar me dy modelet e tjera dhe K-NN ka performancën më të ulët se dy modelet e tjera.

Decision Tree (DT)

Algoritmet DT përdorin paraqitjen si pemë për të mësuar nga të dhëna të etiketuara dhe më pas për të klasifikuar të dhëna të reja. Çdo nyje e brendshme e pemës i korrespondon një atributi dhe çdo gjethe i korrespondon një etikete të klasës. C4.5 është algoritmi DT më popullor, i cili është një klasifikues statistikor që përdor teknikën përça dhe sundo për të mësuar nga të dhënat trajnuese. Algoritmi C4.5 fillon krijimin e pemës duke i ndarë të dhënat në grupe duke përcaktuar vlerën optimale dhe më të mirë për klasifikim. I njëjti proces përsëritet për çdo grup të dhënash derisa algoritmi të përcaktojë që të gjithë të dhënat në një grup i përkasin të njëjtës klasë ose kur grupi është mjaftueshëm i vogël. Në këtë mënyrë çdo nyje e brendshme përmban një kusht, dhe vlera e kushtit përdoret për të përcaktuar degën që do të ndiqet. Të dhënat e reja do të klasifikohen sipas etiketës së gjethes që arrihet. Ky algoritëm ka performancë shumë të mirë në përpunimin e sasive të

mëdha të dhënash. Në Weka J48 është implementimi i algoritmit C4.5 (Witten, et al., 2017).

Wang dhe Gao (2015) kanë përdorur pesë algoritmet NB, SVM, BN, C4.5 dhe Random Forest për të krijuar një model për klasifikimin e tweet-ve të opinioneve për kompanitë ajrore. Modeli bazohet në një strategji votimi ku modelet e ndërtuar veçmas nga secili nga algoritmet kanë të njëjtën peshë. Çdo tweet klasifikohet në mënyrë të pavarur nga çdo model i algoritmeve dhe klasa përfundimtare e tij përcaktohet si klasa që ka më shumë vota nga të pesë modelet e algoritmeve. Rezultatet tregojnë se strategjia e propozuar ka performancë më të mirë se në rastin kur algoritmet përdoren individualisht në krijimin e një modeli.

Në një tjetër punim është propozuar një sistem i ngjashëm votimi duke përdorur tri algoritme të tjerë NB, SVM dhe Bagging për klasifikimin e opinioneve në gjuhën turke si pozitive, negative dhe neutrale. Kjo zgjidhje e propozuar ka performancë më të mirë krahasuar me performancën kur tre klasifikuesit përdoren individualisht (Catal & Nangir, 2017).

Rrjet Neural Artificial (RNA) (angl. Artificial Neural Network (ANN))

Rrjetet neurale artificiale vitet e fundit kanë tërhequr vëmendjen e shumë kërkuesve për t'u përdorur për opinion mining. RNA ka për qëllim të nxjerrë karakteristika duke kombinuar në mënyrë lineare të dhënat hyrëse dhe më pas të modelojë daljen si një funksion jo linear të këtyre karakteristikave. Ato paraqiten si një diagram rrjeti që përfshin nyje të lidhura midis tyre. Nyjet janë të vendosura në shtresa dhe arkitektura tipike e një rrjeti neural përmban tri shtresa, shtresa e hyrjes, shtresa e daljes dhe një shtresë të fshehur (Moraes, et al., 2013).

Të mësuarit e thelluar (MTh) (angl. Deep learning (DL))

Të mësuarit e thelluar është aplikimi i rrjeteve neurale artificiale për të mësuar dhe parashikuar më pas duke përdorur rrjete me shumë shtresa. Avantazhi i tyre është aftësia më e lartë për të mësuar në krahasim me rrjetet neurale që zakonisht kanë një deri në tri shtresa dhe një sasi të vogël të dhënash (Zhang, et al., 2018).

Në punimin e Moraes et al. (2013) propozohet një rrjet neural për klasifikimin e opinioneve sipas polaritetit në dy klasa pozitive dhe negative duke e krahasuar me një model të bazuar në algoritmin SVM. Rezultatet tregojnë që modeli i rrjetit neural performon më mirë se modeli i SVM.

Tang et al. (2015) kanë propozuar përdorimin e një rrjeti neural për ndërtimin e një modeli për klasifikimin e opinioneve me pikë nga 1 deri në 5 duke marrë parasysh edhe se kush e ka shprehur atë opinion. Metoda e propozuar përshin dy komponentë, komponenti i parë modifikon vektorin e fjalës sipas personit që ka shprehur opinionin dhe më pas komponenti i dytë trajnohet për të parashikuar vlerësimin me pikë. Modeli i propozuar është testuar në dy korpuse dhe rezultatet tregojnë që ka performancë të mirë.

Për të eliminuar problemin e mungesës së memorie në punimin e Xu et al. (2016) propozohet përdorimi i një rrjeti neural *cached* LSTM për klasifikimin e opinioneve për të nxjerrë informacione të përgjithshme semantike nga opinione të shprehura në tekste të gjata. Rrjeti i propozuar ka një mekanizëm *cache* që e ndan memorien në 7 grupe me nivele të ndryshme harrese në mënyrë që rrjeti të ruajë më mirë informacionin emocional në një njësi rekurente. Zgjidhja e propozuar është vlerësuar në tri korpuse opinionesh në nivel dokumenti duke rezultuar me performancë të lartë.

RNA dhe MTh janë përdorur jo vetëm në klasifikimin e opinioneve në nivel dokumenti por dhe në nivele të tjera më të detajuara si fjalie dhe aspekti. Në punimin e Poria et al. (2016), autorët kanë propozuar përdorimin për herë të parë DL për të nxjerrë aspekte nga opinionet për Opinion Mining. Ata kanë përdorur një rrjet të thellë CNN me 7 shtresa të kombinuar me një set modelesh gjuhësore për të etiketuar çdo fjalë në fjali si një aspekt ose jo dhe polaritetin e tyre. Zgjidhja e propozuar ka saktësi më të lartë se metodat tradicionale. Kurse në punimin e tyre Pham dhe Le (2017) kanë propozuar një model me arkitekturë me shumë shtresa përfaqësimi të njohurive për të paraqitur nivele të ndryshme të ndjenjave në një opinion tekst. Ky model paraqitjeje integrohet me një rrjet neural për të ndërtuar një model parashikimi për polaritetin e mendimit të opinionit në tërësi. Kjo duke u bazuar që një opinion shpreh ndjenja të ndryshme për aspekte të ndryshme të objektit që i referohet. Rezultatet eksperimentale në një korpus opinionesh të klientëve të hoteleve në TripAdvisor tregojnë që modeli i propozuar ka performancë të mirë.

3.4.1.2. Teknikat e të mësuarit gjysmë të kontrolluar

Etiketimi i të dhënave është një proces i kushtueshëm që kërkon kohë dhe persona që të kenë aftësi të mira etiketimi për ta realizuar. Metodat e të mësuarit gjysmë të kontrolluar përdorin një sasi të vogël të dhënash të etiketuara fillestare dhe një sasi të madhe të dhënash të paetiketuar për të ndërtuar një model klasifikimi. Të dhënat e paetiketuara përdoren për të plotësuar informacionin e marrë nga të dhënat e etiketuara në fazën e trajnimit. Këto teknika janë shumë të përdorshme në rastin kur kemi sasi të mëdha të dhënash të paetiketuara në dispozicion. Në punimin e tyre Silva et al. (2016) kanë realizuar një përmbledhje të teknikave të të mësuarit gjysmë të kontrolluar dhe kanë realizuar një vlerësim eksperimental të tri teknikave të tilla, me vetëtrajnim (angl. self-training), bashkëtrajnim (angl. co-training) dhe modelimi sipas fushës (angl. topic modeling) për klasifikimin e tweet-eve.

Teknika me vetëtrajnim (angl. self-training)

Fillimisht ndërtohet një model duke trajnuar klasifikuesin me sasinë e vogël të të dhënave të etiketuara në dispozicion. Më pas ky model do të përdoret për të etiketuar të dhëna të reja nga korpusi i të dhënave të paetiketuara. Ndër këto të dhëna të reja të etiketuara ato që janë etiketuar me besueshmëri maksimale u shtohen të dhënave fillestare të etiketuara trajnuese. Në këtë mënyrë do të krijohet një korpus i etiketuat më i madh i cili

përdoret për të ritrajnuar klasifikuesin duke krijuar një model të ri. Ky proces ripërsëritet derisa të kemi modelin dhe sasinë e korpusit të etiketuar që na nevojitet. Një problem që haset në përdorimin e kësaj metode është që nëse modeli i trajnuar i ka parashikuar gabim etiketën e të dhënave ato do të modifikojnë modelin në mënyrë të gabuar. Për të shmangur këtë problem, Hong et al. (2014) kanë propozuar një metodë konkurruese me vetëtrajnim. Autorët kanë krijuar tre modele mbi bazën e tri perspektivave *threshold*, numër i njëjtë dhe numri më i madh i rifreskimeve dhe më pas kanë zgjedhur modelin që ka *F-measure* më të lartë. Dhe në këtë mënyrë arrihet të përmirësohet performanca e modelit të klasifikimit të opinioneve. Në punimin e tyre Asch dhe Daelemans (2016) shprehen se ngjashmëria midis korpuseve mund të përdoret për identifikimin e karakteristikave nga të cilat metodat me vetëtrajnim mund të përfitojnë.

Teknikat me bashkëtrajnim (angl. co-training)

Në ndryshim nga teknikat me vetëtrajnim në teknikat me bashkëtrajnim korpusi i vogël i etiketuar përdoret për të trajnuar dy klasifikues të ndryshëm për aspekte të ndryshme të korpusit së trajnimit. Dy modelet e trajnuara përdoren për të etiketuar korpusin e paetiketuar. Të dhënat e reja të etiketuara nga njëri nga modelet përdoren për të ritrajnuar modelin tjetër. Në këtë mënyrë gjenerohet një korpus më i madh i etiketuar dhe modelet përmirësohen duke mësuar nga një sasi më e madhe të dhënash.

Carter dhe Inkepen (2015) në punimin e tyre kanë propozuar një algoritëm me bashkëtrajnim për identifikimin e aspekteve dhe polaritetin e mendimit të shprehur për to në opinione për produkte. Vlerësimi krahasues i këtij propozimi me metodat ekzistuese në dy korpuse opinionesh për laptop dhe restorante tregojnë se algoritmi ka saktësi më të lartë se metodat ekzistuese.

Rezultatet e vlerësimit eksperimental në punimin e Hong et al. (2014) tregojnë që metoda me bashkëtrajnim ka performancë më të mirë kur opinionet nuk përmbajnë sarkazëm dhe ironi dhe sasia e të dhënave është e limituar. Kurse metoda me vetëtrajnim është zgjedhja më e mirë kur kemi një sasi relativisht të madhe të dhënash të etiketuara.

Në një tjetër punim autorët kanë krahasuar performancën e metodave me vetëtrajnim dhe me bashkëtrajnim për klasifikimin e tweet-eve sipas polaritetit të mendimit të shprehur. Dy metodat janë përdorur për të krijuar një korpus më të zgjeruar me tweet-e duke u nisur nga një korpus relativisht i vogël tweet-esh të etiketuara. Rezultatet eksperimentale tregojnë që metoda me bashkëtrajnim ka performanë më të lartë kur kemi numër të limituar etiketash kurse metoda me vetëtrajnim ka performancë më të lartë kur kemi sasi të mëdha të dhënash të etiketuara (Iosifidis & Ntutsi, 2017).

Teknikat e të mësuarit gjysmë të kontrolluar kanë një përdorim të gjerë në *cross-lingual* OM. Në punime të ndryshme janë studiuar dhe propozuar teknika të ndryshme gjysmë të kontrolluara për këtë çështje. Vlerësimet e ndryshme eksperimentale tregojnë që këto metoda kanë performancë të mirë për OM në gjuhë të ndryshme. Në dy punimet e tyre Hajmohammadi et al. (2014; 2015) kanë propozuar një metodë të të mësuarit nga shumë

pamje (angl. multi view learning) dhe një metodë të bazuar në grafe (angl. graph-based) në për cross-lingual Opinion Mining. Në Hajmohammadi et al. (2014), autorët kanë propozuar një model të të mësuarit gjysmë të kontrolluar që përdor të dhëna të etiketuara në gjuhë të ndryshme për të përfshirë të dhëna të paetiketuara në një gjuhë tjetër në procesin e mësimit. Kjo metodë është përdorur për të shmangur problemet që shfaqen kur përdoren aplikacione të përkthimit automatik në klasifikimin e opinioneve në gjuhë të ndryshme në dy klasa si pozitive dhe negative, si kuptimi i ndryshëm i shprehjeve në gjuhë të ndryshme apo pamundësia për të përkthyer çdo term nga një gjuhë në tjetrën. Metoda është vlerësuar në një korpus opinionesh librash në pesë gjuhë, anglisht, frëngjisht, japonisht dhe në gjuhën kineze. Rezultatet tregojnë se metoda e propozuar ka performancë shumë të lartë për klasifikimin e opinioneve në cross-lingual. Kurse, në Hajmohammadi et al. (2015) është propozuar një metodë e bazuar në grafë për klasifikimin e opinioneve në dy klasa si pozitive dhe negative për cross-lingual. Metodat e bazuara në grafe i paraqesin të dhënat si grafe të peshuar në të cilën nyjet përfaqësojnë instancat dhe lidhjet paraqesin ngjashmërinë e instancave. Në këto metoda supozohet që instancat që kanë lidhje më të forta midis tyre i përkasin të njëjtës klasë. Ne rastin e klasifikimit të opinioneve instancat janë dokumente që përmbajnë opinionin. Metoda e propozuar përdor një sistem automatik përkthimi për të përkthyer opinionet nga një gjuhë në tjetrën. Duke u bazuar në ngjashmërinë midis të dhënave të etiketuara dhe atyre të paetiketuar janë ndërtuar respektivisht dy grafe. Klasa që i përket një opinioni të paetiketuar përcaktohet mbi bazën e rezultateve të të dhënave të etiketuara dhe të paetiketuara. Rezultatet eksperimentale tregojnë që metoda e propozuar ka performancë më të mirë se metodat tradicionale, me bashkëtrajnim, SCL, dhe SVM.

Ange et al. (2018) në punimin e tyre kanë propozuar një metodë të të mësuari gjysmë të kontrolluar për klasifikimin e opinioneve në tri klasa, pozitive, negative dhe neutrale që përdor një arkitekturë multimodale e cila përdor avantazhet e të mësuarit e thelluar dhe të të dhënave për të marrë apo nxjerrë informacion më të detajuar. Modeli i propozuar merr në konsideratë natyrën multimodale të sjelljeve njerëzore në procesin e klasifikimit. Modeli i propozuar është një kombinim i nje rrjeti LSTM, një auto-enkoderi LSTM, një rrjeti CNN dhe rrjeteve të shumëfishtë DNN (Deep Neural Network). Modeli i propozuar është vlerësuar eksperimentalisht dhe krahasuar me performancën e algoritmeve ekzistues për të njëjtën detyrë. Rezultatet tregojnë që modeli i propozuar ka performancë më të mirë krahasuar me teknikat tradicionale.

3.4.1.3. Teknikat e të mësuarit e pakontrolluar

Teknikat e të mësuarit e pakontrolluar (angl. unsupervised learning) mësojnë nga të dhëna të paetiketuara dhe përdoren për të zbuluar karakteristika të fshehura në këto të dhëna të paetiketuara. Në problemin e klasifikimit të tekstit dhe veçanërisht në klasifikimin e opinioneve këto teknika kanë aftësinë që të realizojnë detyrën e caktuar pa patur nevojë për njohuri gjuhësore ose për t'u trajnuar me të dhëna të etiketuara (Russell & Norvig, 2020). Gjithnjë e më shumë kërkuesit po përqendrohen në zhvillimin e teknikave të tilla

qoftë për detyrën e klasifikimit të opinioneve sipas mendimit të shprehur qoftë për detyra të tjera të Opinion Mining.

Grupimi (angl. Clustering) është një nga teknikat më të përdorshme të të mësuarit të pakontrolluar. Grupimi ka për qëllim të identifikojë grupe të dhënash në të dhënat e paetiketuara kur nuk ka as klasa të përcaktuara. Të dhënat grupohen në grupe sipas ngjashmërive për një karakteristikë të caktuar. Të dhënat që grupohen në një grup të caktuar kanë ngjashmëri maksimale dhe të dhënat e grupeve të ndryshme kanë ngjashmëri minimale (Russell & Norvig, 2020).

K-means është një algoritëm grupimi nga më të thjeshtët. Ky algoritëm i grupon të dhënat duke i klasifikuar në një numër apriori k grupesh (angl. clusters). Në këtë mënyrë qëllimi kryesor i këtij algoritmi është të përcaktojë k qendra, një për çdo grup, të cilat janë të vendosura në mënyrë dinamike sepse vendndodhje të ndryshme kanë rezultate të ndryshme. Çdo pikë që përcakton një grup të dhënash i bashkëngjitet qendrës më të afërt duke realizuar një grupim paraprak. Më pas rillogariten k qendra të reja si qendra të mbyllura. Dhe në fund realizohet lidhja e pikave ekzistuese me qendrat e reja. Kjo procedurë do të realizohet në mënyrë ciklike duke realizuar ndryshimin hap pas hapi të qendrave deri sa ato të mos ndryshojnë më. Li dhe Liu (2012) propozuan për herë të parë përdorimin e këtij algoritmi duke e kombinuar me një metodë TF-IDF të peshuar dhe një metodë votimi për rezultat sa më të qëndrueshëm për klasifikimin e opinioneve sipas polaritetit të mendimit të shprehur. Modeli i propozuar është vlerësuar eksperimentalisht duke përdorur një korpus opinionesh për filma të klasifikuar në dy klasa, pozitive dhe negative. Rezultatet eksperimentale tregojnë se propozimi ka performancë të mirë edhe pse më e ulët krahasuar me teknikat e të mësuarit të kontrolluar.

Në punimin e Unnisa et al. (2016) propozohet një metodë grupimi spektrale (angl. spectral clustering) duke përdorur algoritmin k-means për klasifikimin e tweets-ve në dy klasa, pozitive dhe negative. Rezultatet eksperimentale tregojnë se metoda e propozuar ka performancë më të mirë se SVM, ME dhe NB dhe se kjo metodë nuk është specifike vetëm për opinione për filma por mund të aplikohet lehtësisht dhe në fusha të tjera.

Dy nga problemet e teknikave që paraqesin të dhënat si vektor me gjatësi të caktuar, si *bag-of-words*, në analizimin e opinioneve janë humbja e renditjes së fjalëve dhe mos marrja parasysh e semantikës së fjalëve. Për të anashkaluar këto probleme, Le dhe Mikolov (2014) kanë propozuar një algoritëm të të mësuarit e pakontrolluar, Paragraph Vector, që mëson paraqitjen e të dhënave me gjatësi të caktuar nga të dhëna me gjatësi variabël si fjali, paragraf apo dokument. Rezultatet eksperimentale në dy korpuse opinionesh tregojnë që algoritmi i propozuar ka performancë më të mirë se modelet tradicionale dhe ka aftësinë të kapërcejë probleme që shfaqen në modelin *bag-of-words*.

Teknikat e të mësuarit e pakontrolluar janë përdorur dhe për identifikimin e aspekteve dhe klasifikimin e opinioneve sipas polaritetit të mendimeve të aspekteve të identifikuar. Federici dhe Dragoni (2017), në punimin e tyre, kanë propozuar një metodë të të mësuarit e pakontrolluar që ka si qëllim identifikimin e aspekteve dhe polaritetit të mendimit të

shprehur lidhur me këto aspekte në një opinion tekst. Kjo metodë ka performancë të mirë në rastin e korpusit të opinioneve për laptopë kurse në korpusin e opinioneve për restorantet renditet i treti. Autorët kanë argumentuar këtë rezultat me faktin që gjuha e përdorur në opinionet për laptopë është më e thjeshtë se ajo për restorantet. Një nga problemet në metodat e të mësuarit e pakontrolluar për identifikimin e aspekteve është identifikimi i aspekteve false pozitive.

Për të përfituar nga avantazhet e metodave të të mësuarit e pakontrolluar si mos nevoja e një korpusi të etiketuar dhe adaptimi në fusha të ndryshme opinionesh Vilares et al. (2017) propozojnë një metodë të tillë për analizimin e opinioneve në shumë gjuhë. Metoda e propozuar bazohet në një grup rregullash sintaksore për përcaktimin e marrëdhënieve semantike midis fjalëve dhe në konceptin e veprimeve të përbëra. Rezultatet eksperimentale duke përdorur tri korpuse të ndryshme opinionesh të klasifikuara në dy klasa, pozitive dhe negative, tregojnë që metoda e propozuar ka performancë të lartë dhe operacionet e përbëra mund të përdoren për gjuhë të ndryshme.

3.4.2. Teknika të bazuara në leksik

Teknikat e bazuara në leksik konsiderohen si teknika të të mësuarit e pakontrolluar të cilat bazohen në një leksik ndjenjash për të realizuar klasifikimin e opinioneve sipas polaritetit të ndjenjës së shprehur. Fjalë të caktuara të përdorura në një opinion mund të shprehin ndjenja, si për shembull fjalë si: mirë dhe bukur shprehin një ndjenjë pozitive kurse fjalë si: keq dhe shpifur shprehin një ndjenjë negative. Në detyrën e analizimit të opinioneve mund të merret parasysh polariteti i ndjenjës që shpreh një fjalë e veçantë apo një shprehje për të përcaktuar polaritetin e mendimit të opinionit. Fjalët ose shprehjet që shprehin një ndjenjë mund të jenë të thjeshta siç kemi shprehur më sipër ose shprehje krahasuese si: më mirë se, më keq se, më i bukur se, etj. Në gjuhën shqipe zakonisht fjalët që shprehin ndjenja janë mbiemra ose ndajfolje por gjithashtu dhe emra si dashuri, urrejtje, paqe, etj. shprehin ndjenja. Në metodat e bazuara në leksik për të realizuar klasifikimin e opinioneve sipas polaritetit të mendimit të shprehur përdoret një fjalor ku çdo fjalë apo shprehje e cila shpreh një ndjenjë shoqërohet me polaritetin e ndjenjës që shpreh dhe nuk është e nevojshme të kemi një korpus trajnimi të etiketuar. Në këtë fjalor polariteti i përcaktuar mund të jetë i peshuar ose jo, ku pesha përcaktohet nëpërmjet një numri që përfaqëson nivelin e pozitivitetit apo negativitetit të ndjenjës së shprehur nga fjala ose shprehja, duke formuar në këtë rast fjalorin e peshuar apo të papeshuar. Teknikat e bazuara në leksik për krijimin e fjalorëve mund të klasifikohen në tre grupe: manuale, të bazuara në fjalor gjuhësor (angl. dictionary-based approach) ose në korpus (angl. corpus-based approach). Teknikat manuale kërkojnë aftësi të mira njerëzore për t'u zhvilluar dhe gjithashtu kohë. Kjo për arsye që etiketimi i polaritetit të ndjenjës së shprehur nga fjala bëhet manualisht. Për këto arsye këto teknika manuale përdoren kryesisht për kontrollin e saktësisë së fjalorëve që janë krijuar automatikisht nëpërmjet teknikave të bazuara në fjalor ose korpus. Në përdorimin e teknikave të bazuara në leksik në klasifikimin e opinioneve

dy janë problemet që duhen adresuar: si të identifikohen fjalët apo shprehjet që shprehin ndjenja dhe si të identifikohen dhe veprohet me fjalët ose shprehjet të cilat janë të varura nga fusha/objekti për të cilin shprehet opinioni (Liu, 2015).

Manuale

Në punimin e tyre Khoo dhe Johnkhan (2018) kanë propozuar një fjalor ndjenjash, WKWSCI, që është përdorur në një metodë të bazuar në leksik për kategorizimin e opinion në nivel dokumenti dhe fjalie sipas polaritetit të mendimit të shprehur. Ky fjalor është ndërtuar në mënyrë manuale në dy faza. Në fazën e parë fjalët u klasifikuan si pozitive, negative ose neutrale dhe në fazën e dytë fjalët pozitive u klasifikuan në tre nivele pozitiviteti, pak, neutrale, shumë pozitive. Fjalori i propozuar është krahasuar me 5 fjalorë të tjerë ekzistues, Hu & Liu Opinion Lexicon, Multi-perspective Question Answering (MPQA) Subjectivity Lexicon, General Inquirer, National Research Council Canada (NRC), Word-Sentiment Association Lexicon dhe Semantic Orientation Calculator (SO-CAL) duke përdorur një korpus opinionesh për produkte të marrë nga Amazon dhe një korpus lajmesh. Fjalori i propozuar është vlerësuar eksperiemntalisht sipas disa skenarëve: një modeli të bazuar në leksik të thjeshtë dhe një modeli të bazuar në leksik që përdor regresionin logjik për të përcaktuar peshën e kategorive të ndryshme të fjalëve për klasifikimin në nivel dokumenti dhe fjalie. Rezultatet eksperimentale tregojnë se metodat e bazuar në leksikon kanë performancë të mirë në klasifikimin e opinioneve në nivel dokumenti dhe fjalie sipas polaritetit të mendimit të shprehur. Fjalorët WKWSCI, MPQA, Hu & Liu dhe SO-CAL kanë performancë të njëjtë në klasifikimin e opinioneve të produkteve. Fjalori i propozuar, WKWSCI, ka performancë të mirë në klasifikimin e opinioneve të lajmeve kurse fjalori i Hu & Liu Opinion lexicon ka performancë më të mirë kur përdoret për klasifikimin e opinioneve të produkteve.

Teknikat e bazuara në fjalor

Në shumë punime janë propozuar metoda të bazuara në fjalor për krijimin e një fjalori më të gjerë emocionesh duke u nisur nga një fjalor me pak fjalë fillestar dhe një algoritëm për zgjerimin automatik të tij. Hu dhe Liu (2004) kanë propozuar një metodë të thjeshtë për të ndërtuar një fjalor fjalësh (mbiemrash) që shprehin ndjenja. Fillimisht ata kanë krijuar manualisht një fjalor të vogël me 30 mbiemra duke përcaktuar polaritetin negativ apo pozitiv të fjalëve. Më pas algoritmi i propozuar e zgjeron këtë fjalor duke kërkuar në fjalorin WordNet për sinonimet dhe antonimet e mbiemrave të fjalorit manual. Mbiemrat e gjeturë i shtohen listës dhe më pas kemi një interacion të dytë të ekzekutimit të algoritmit për të gjetur mbiemra të rinj duke u bazuar në këtë listë më të zgjeruar. Ky proces vazhdon derisa nuk gjenden më mbiemra të rinj. Autorët kanë vlerësuar eksperimentalisht propozimin e tyre në klasifikimin e opinioneve për pesë produkte duke arritur një saktësi në parashikim mesatarisht 84%.

Në punimin e tyre Wesgate dhe Valova (2018) kanë propozuar një algoritëm për të ndërtuar një fjalor duke u bazuar në një graf dhe sinonimet e fjalëve. Algoritmi e përcakton polaritetin e fjalës nëpërmjet tre hapave: gjenerimi i grafit i cili në mënyrë rekursive lidh fjalët me sinonimet e tyre, gjenerimi i rrugëve që lidhin fjalët dhe gjetja e rrugës optimale vlera e të cilit përcakton polariteti e fjalës. Rezultatet eksperimentale tregojnë që kjo metodë ka performancë të mirë në përcaktimin e polariteteve të fjalëve. Autorët kanë përcaktuar se numri ideal i sinonimeve të marra për një fjalë është tetë dhe performanca e algoritmit është e lartë deri në këtë numër në term të përcaktimit të polaritetit të fjalës. Në rastin kur merren tetë deri në dhjetë sinonime për një fjalë performanca e algoritmit bie në përcaktimin e polaritetit kjo për arsye sepse fjalori i sinonimeve i marrë në konsideratë i ka të renditura sinonimet e fjalëve sipas një rendi zbritës të kuptimit dhe nivelit të sinonimisë.

Park dhe Kim (2016) kanë propozuar një fjalor të zgjeruar fjalësh shoqëruar me polaritetin e ndjenjës së tyre. Për të krijuar fjalorin e zgjeruar metoda e propozuar përdor një fjalor me fjalë bazë të etiketuara sipas polariteti të tyre dhe tre fjalorë tradicional online. Fillimisht nga secili nga fjalorët online identifikohen sinonimet dhe antonimet për fjalët e fjalorit bazë. Më tej në fjalorin bazë shtohen vetëm sinonimet ose antonimet të cilat ndodhen në të tre fjalorët për një fjalë. Këto veprime ripërsëriten për të krijuar një fjalor sa më të pasur. Në këtë mënyrë është krijuar një fjalor i gjerë duke u nisur nga një fjalor bazë i etiketuar me pak fjalë dhe pa nevojën e burimeve njerëzore për etiketim. Sa më i gjerë të jetë fjalori aq më rezultate të sakta do kemi në klasifikimin e opinioneve sipas polaritetit të mendimit të shprehur. Për të vlerësuar eksperimetalisht propozimin e bërë është marrë në konsideratë një korpus me tweet-e të klasifikuara në pozitive, negative dhe neutrale. Rezultatet tregojnë që përdorimi i këtij fjalori më të zgjeruar është efektiv në klasifikimin e tweet-eve.

Teknikat e bazuara në korpus

Kuptimi dhe polariteti i një fjale mund të ndryshojë nga një fushë në tjetrën. Një fjalë mund të ketë polaritet pozitiv në një kontekst të caktuar dhe në një tjetër të ketë polaritet negativ. Për këtë arsye dhe përcaktimi i polaritetit të një opinioni do të varet nga fusha e opinionit, prandaj shpesh herë mund të jetë e nevojshme të krijohen fjalorë duke u bazuar në fushën e opinionit ku do përdoret. Metodat e bazuara në korpus janë zhvilluar në dy skenarë kryesorë: duke u nisur nga një listë fjalësh të etiketuara që shprehin ndjenja identifikohen fjalë të tjera dhe orientimi i tyre nga një korpus i një fushe të caktuar dhe përshtatja e një fjalori që mban fjalë të përgjithshme që shprehin ndjenja në një fjalor të ri duke përdorur një korpus opinionesh të një fushe të caktuar. Në një korpus fjalët që shprehin mendim dhe polariteti i tyre mund të identifikohen duke shfrytëzuar rregulla gjuhësore për lidhëzat ose duke përdorur marrëdhëniet sintaksore që ekzistojnë midis opinioneve dhe aspekteve (Liu, 2015).

Molina-González et al. (2015) kanë krijuar fjalorin iSOL që përmban informacion edhe rreth fushës së opinioneve në spanjisht. Procesi i krijimit të fjalorit ka nisur me përkthimin

nga anglishtja në spanjisht të fjalorit BLEL. Më pas ky fjalor u pasurua në dy mënyra me fjalë nga tetë fushat e opinioneve të marra në konsideratë. Mënyra e parë konsiston në identifikimin dhe shtimin e fjalëve specifike për çdo fushë dhe mënyra e dytë konsiston në identifikimin dhe shtimin e fjalëve që mund të mos shprehin një ndjenjë por që janë më të përdorurat në korpus. Rezultatet eksperimentale tregojnë se metoda e propozuar është efektive.

Chiavetta et al. (2016) kanë propozuar një metodë të bazuar në korpus për klasifikimin e opinioneve për libra në italisht. Fjalori në gjuhën italiane për fjalët dhe polaritetin e tyre është krijuar në tri faza: në fazën e parë çdo fjalë e korpusit kontrollohet nëse gjendet ose jo në dy fjalorët MultiWordNet dhe SentiWordNet, nëse gjendet shtohet në fjalor nëse nuk gjendet shtohet në një fjalor tjetër të quajtur fjalët që mungojnë. Në fazën e dytë etiketohen manualisht dhe shtohen në fjalor fjalët nga fjalori i fjalëve që mungojnë. Në fazën e tretë autorët kanë shtuar në fjalor fjalë që shprehin një ndjenjë dhe që lidhen specifikisht me fushën e opinioneve. Polariteti i fjalisë përcaktohet si shumë e polaritetit të çdo fjale. Dhe më pas polariteti i opinionit në tërësi (si dokument) përcaktohet si shuma e polaritetit të çdo fjalie. Sistemi i propozuar ka saktësi mbi 82% në klasifikimin e opinioneve në gjuhën italiane nga Amazon në dy klasa pozitive dhe negative.

3.4.3. Teknika hibride

Në shumë punime kërkues të ndryshëm kanë përdorur së bashku metodat e të mësuarit e automatizuar dhe metodat e bazuara në fjalor për të ndërtuar modele më komplekse dhe më eficiente në klasifikimin e opinioneve në nivel dokumenti sipas polaritetit të mendimit të shprehur.

Gautam dhe Yadav (2014) kanë vlerësuar performancën e tri algoritmeve të të mësuarit të kontrolluar, ME, NB dhe SVM duke i kombinuar me fjalorin WordNet për klasifikimin e tweet-ve si pozitive dhe negative. Rezultatet tregojnë që algoritmi NB ka performancën më të mirë dhe përdorimi i WordNet sjell përmirësim të performancës së algoritmave. Modeli hibrid që përdor fjalorin e fjalëve që shprehin një ndjenjë dhe një algoritëm të të mësuarit e automatizuar ka performancë më të mirë se në rastin e përdorimit vetëm të algoritmit të të mësuarit e automatizuar. Kjo sepse nga kombinimi i realizuar përfitojmë nga përparësitë e të dyja metodave të përdorura.

Teng et al. (2016) kanë propozuar një metodë hibride e cila merr në konsideratë dhe kontekstin në të cilën përdoret një fjalë. Modeli i propozuar bazohet në një rrjet neural, BiLSTM, për të mësuar fuqinë, intensitetin dhe mohimin e ndjenjës të fjalëve që përbëjnë opinionin. Autorët kanë përdorur dhe vlerësuar performancën e metodës së propozuar duke përdorur katër fjalorë të ndryshëm TS-Lex, S140-Lex, SD-Lex dhe SWN-Lex në tri korpuse opinionesh të klasifikuara si pozitive dhe negative. Rezultatet eksperimentale tregojnë që metoda hibride e propozuar ka performancë më të mirë se në rastin e përdorimit të një modeli të bazuar vetëm të rrjeteve neurale.

Ndërkohë, Al-Sharuee et al. (2018) kanë propozuar një metodë hibride të të mësuarit e pakontrolluar që kombinon teknikat e të mësuarit e automatizuar dhe të bazuara në leksik për të zgjidhur problemin e varësisë nga fusha e opinioneve dhe kostos për krijimin e korpusit të etiketuar. Metoda e propozuar përdor fjalorin SentiWordNet dhe algoritmin K-means që i përcakton qendrat në mënyrë jo-rastësore. Rezultatet e eksperimenteve në shtatë korpuse opinionesh nga fusha të ndryshme të kategorizuara si pozitive dhe negative tregojnë se kjo metodë ka performancë më të lartë se metodat tradicionale.

3.5. Kriteret e vlerësimit

Parametrat për vlerësimin e teknikave MA për detyrën e klasifikimit të opinioneve sipas polaritetit të mendimit të shprehur, në Opinion Mining janë adaptuar nga fusha tradicionale e klasifikimit të teksteve (Webb, 2011). Këto parametra janë: saktësia (angl. accuracy), precizoni (angl. precision), recall dhe F-measure (Ting, 2011).

Le të supozojmë se duam të vlerësojmë performancën e një modeli MA për të përcaktuar nëse një opinion për një entitet të caktuar është pozitiv apo negativ. Modeli MA i trajnuar nga njohuritë që ka mësuar duhet të parashikojë nëse opinioni është pozitiv apo negativ.

Në Tabelën 3.1 paraqitet matrica e saktësisë së klasifikimit të opinioneve.

Tabela 3.1 Matrica e saktësisë së klasifikimit të opinioneve

		Klasifikimi nga modeli MA	
		Pozitiv	Negativ
Klasifikimi real	Pozitiv	PP	PN
	Negativ	NP	NN

Nga parashikimet e modelit marrim të dhëna për parametrat:

- PP – është numri i opinioneve pozitive që janë klasifikuar saktë si pozitive dhe nga modeli;
- PN – është numri i opinioneve pozitive që janë klasifikuar gabim nga modeli si opinione negative;
- NP – është numri i opinioneve negative që janë klasifikuar gabim nga modeli si opinione pozitive;
- NN – është numri i opinioneve negative që janë klasifikuar saktë si negative dhe nga modeli.

Saktësia përcakton se sa herë modeli ka realizuar parashikim të saktë. Llogaritet si raport i numrit të opinioneve të klasifikuara siç duhet nga modeli me numrin e të gjithë opinioneve të klasifikuar. Formula e llogaritjes është:

$$Saktësia = \frac{PP + NN}{PP + PN + NP + NN}$$

Saktësia lidhet drejtpërdrejt me nivelin e gabimit (angl, error rate) që është madhësia që përdoret për të matur nivelin e gabimit të parashikimit të modelit kundrejt klasifikimit real. Lidhja midis këtyre dy madhësive është:

$$saktësia = 1.00 - \text{gabimi}$$

Precizioni përcakton sa herë modeli ka parashikuar saktë që një opinion është pozitiv. Llogaritet si raporti i numrit të opinioneve të cilat janë klasifikuar saktë nga modeli si opinione pozitive me numrin total të opinioneve që janë klasifikuar si pozitive nga modeli. Formula e llogaritjes është:

$$Precizioni = \frac{PP}{PP + NP}$$

Recall përcakton sa nga opinionet pozitive janë klasifikuar saktë si opinione pozitive nga modeli. Llogaritet si raporti i numrit të opinioneve pozitive që janë klasifikuar saktë si pozitive nga modeli me numrin total të opinioneve pozitive (opinionet pozitive të klasifikuara si pozitive nga modeli dhe opinionet pozitive të klasifikuara si negative nga modeli). Formula e llogaritjes është:

$$Recall = \frac{PP}{PP + PN}$$

F-measure llogaritet duke kombinuar vlerat e precizonit dhe recall sipas formulës:

$$F - measure = 2 * \frac{precizon * recall}{precizon + recall}$$

Vlera e madhësive të diskutuar më lartë është në diapazonin e vlerave nga 0 në 1 ose nëse shprehen në përqindje nga 0 në 100%.

KREU 4

GRAMATIKA E GJUHËS SHQIPE

Në këtë kapitull realizohet një vështrim i përgjithshëm i gramatikës së gjuhës shqipe. Gjuha shqipe është një degë e veçantë e familjes së gjuhëve Indo-Evropiane dhe nuk ka ndonjë lidhje prejardhjeje me asnjërën prej gjuhëve të kësaj familje. Për të realizuar përmbledhjen e karakteristikave gramatikore të gjuhës shqipe jemi bazuar në librat "Gramatika e Gjuhës Shqipe 1" (Agalliu, et al., 2002) dhe "Gramatika e Gjuhës Shqipe 2" (Çeliku, et al., 2002).

4.1. Morfologjia e gjuhës shqipe

Morfologjia është ajo pjesë e gramatikës e cila studion tipet kryesore të fjalëformimit, tipi morfologjik dhe atë morfologjiko-sintaksor dhe trajtëformimit. Pra morfologjia merret me studimin e formave që merr një fjalë e lakueshme gjatë lakimit dhe një fjalë e zgjedhueshme gjatë zgjedhimit. Në ndryshim nga leksikologjia që studion fjalën si njësi të fjalorit, morfologjia ka si objekt të studiojë fjalën si një pjesë e ligjëratës duke marrë në konsideratë format dhe kuptimet gramatikore. Pjesët e ligjëratës në gjuhën shqipe janë: emri, mbiemri, numërori, përemri, folja, ndajfolja, parafjala, lidhëza, pjesëza dhe pasthirrma. Këto pjesë të ligjëratës ndahen në të ndryshueshme dhe në të pandryshueshme. Emri, mbiemri, numërori, përemri, folja dhe ndajfolja janë pjesë të ndryshueshme ligjëratës, që do të thotë ato ndryshojnë formë gjatë përdorimit në fjali. Kurse parafjala, lidhëza, pjesëza dhe pasthirrma janë pjesë të pandryshueshme të ligjëratës dhe nuk ndryshojnë formë gjatë përdorimit në fjali.

4.1.1. Emri

Emri është pjesë e ndryshueshme e ligjëratës, që emërton një qenie të gjallë apo një send. Nga ana leksiko-gramatikore emri mund të jetë i përgjithshëm ose i përveçëm, konkret ose abstrakt, përmbledhës, dhe i lëndës. Emri ka karakteristikat gramatikore të trajtës, rasës, numrit dhe gjinisë.

Trajta

Kemi dy trajta të emrit, trajtë e shquar dhe trajtë e pashquar. Emrat në trajtën e pashquar marrin përpara fjalën "një ", "ca" ose "disa" në varësi nëse emri është në numrin njëjës apo shumës.

Rasa

Në varësi të funksionit sintaksor që një fjalë ka në një fjali ndryshon dhe forma e tij. Rasa përcakton këto forma të emrit dhe shpreh lidhjen që ai ka me fjalë të tjera në fjali. Rasat në të cilat lakohet emrin janë pesë: rasa emërore, rasa gjinore, rasa dhanore, rasa kallëzore dhe rasa rrjedhore. Emrat në rasën emërore kanë rolin e kryefjalës. Emrat në rasën gjinore tregojnë përkatësi, për të treguar cilësi apo sendin që mbart një cilësi, personin apo sendin mbi të cilin bie një veprim, për të treguar lidhje, etj.. Në këtë rasë emri shoqërohet gjithmonë me një nga pjesëza: i, e, të, së. Emrat në rasën dhanore zakonisht kanë rolin e kundrinorit të zhdrejtë dhe përdoren vetëm me folje. Emrat në rasën kallëzore zakonisht kanë rolin e kundrinorit të drejtë dhe përdoren vetëm me folje. Emrat në rasën rrjedhore shprehin shkakun, mjetin, cilësinë apo prejardhjen. Zakonisht emri në këtë rasë përdoret me parafjalë si: prej, pas, mbas, prapa, etj.

Numri

Numri përcakton nëse emri përcakton në ose më shumë sende. Në gjuhë shqipe kemi dy numra: njëjës dhe shumës. Një emër mund të marrë të dy numrat, njëjës kur tregon një frymor apo send dhe shumës kur tregon dy ose më shumë frymorë apo sende. Emra që mund të ndodhen në të dyja numrat tregojnë frymorë apo sende të numërueshme. Ka emra që mund të përdoren vetëm në njëjës si dhallë, kollë, dëborë dhe emra që përdoren vetëm në shumës si pantallona, alpe, ethet, etj.

Gjinia

Emrat ndahen në tri gjini: femërore, mashkullore dhe asnjanëse. Një numër shumë i vogël emrash kanë gjini asnjanëse dhe pjesa më e madhe e emrave ka gjininë mashkullore. Në gjuhën shqipe gjinia e emrit shfaqet në të dy numrat, si njëjës dhe shumë. Ekzistojnë disa emra të cilët kanë gjini të ndryshme në këto dy numra dhe emërtohen si emra ambigjenë. Pra, emra që ndryshojnë gjini nga numri njëjës në numrin shumës.

4.1.2. Mbiemri

Mbiemri është pjesë e ndryshueshme e ligjëratës që shpreh një cilësi apo karakteristikë të një frymori, sendi apo dukuri për emrin me të cilin lidhet. Mbiemri përshtate në gjini, numër dhe rasë me emrin që përcakton. Në rastin kur mbiemri përdoret si përcaktor për disa emra, atëherë do të përshtatet në gjini, numër dhe rasë me emrin që ka më pranë.

Në gjuhën shqipe mbiemrat në varësi se si janë formuar ndahen në të nyjshëm dhe të panyjshëm.

Mbiemrat kanë dhe një tjetër kategori gramatikore përveç gjinisë, numrit dhe rasës që është shkalla. Shkalla tregon nivelin e cilësisë së emrit të përcaktuar nga mbiemrit. Mbiemri ka tri shkallë:

1. Shkalla pohore: tregon thjesht cilësinë e frymorit, sendit apo dukurisë pa shprehur ndonjë krahasim, p.sh. i ëmbël, i bukur;

2. Shkalla krahasore: tregon që cilësia e shprehur nga mbiemri krahasohet me veten apo me një cilësi tjetër, p.sh. më i ëmbël se, më i bukur se;
3. Shkalla sipërore: tregon që cilësia e shprehur nga mbiemri është në shkallën më të lartë dhe nuk krahasohet me asnjë cilësi tjetër, p.sh. mjaft i ëmbël, shumë i bukur.

4.1.3. Numërori

Numërori është pjesë e ndryshueshme e ligjëratës, që tregon numra abstraktë a një sasi të caktuar a një pjesë të një grupi qeniesh a sendesh. Numëroret mund të jenë të mirëfilltë që tregon një sasi të plotë ose thyesor që tregojnë një pjesë të një grupi. Duhet theksuar se fjalët të cilat tregojnë radhën e sendeve në gjuhën shqipe janë mbiemra dhe jo numëror, të tillë si : i parë, i dytë, i tretë, etj. Numërori përdoret për të treguar:

- Datë;
- Periudhë të caktuar;
- Sasi të caktuar ose të pacaktuar;
- Përmasa.

Numërorët në përgjithësi nuk kanë kategori gramatikore. Përjashtim bëjnë numërorët tre dhe dy i shoqëruar nga të, që shfaqin kategorinë gramatikore të gjinisë. Numërori tre, në gjininë mashkullore ka trajtën tre dhe në gjininë femërore ka trajtën tri. Numërori dy kur shoqërohet nga të ka trajtën të dy në gjininë mashkullore dhe të dyja në gjininë femërore. Një përjashtim tjetër është kur numërorët përdoren si tregues të emrave dhe marrim si kategorinë gramatikore të gjinisë dhe rrasës.

Thyesat janë shprehje të përbëra nga një numër i plotë dhe një mbiemër prejnumëror i emërzuar në gjininë femërore. Thyesat përdoren vetëm në gjininë femërore. Në rastin kur pjesa e parë e thyesës është fjala një atëherë pjesa e dytë është një mbiemër prejnumëror i emërzuar në numrin njëjës dhe gjininë femërore kurse kur pjesa e parë është dy e më lartë atëherë pjesa e dytë është një mbiemër prejnumëror i emërzuar në numrin shumës dhe gjininë femërore dhe lakohen duke u bazuar në këto karakteristika.

4.1.4. Përemri

Përemri është pjesë e ndryshueshme e ligjëratës e cila tregon në mënyrë të përgjithshme një frymor, send, sasi ose tipar pa e emërtuar atë. Përemri përdoret për të zëvendësuar emrat, mbiemrat ose dhe në raste të veçanta numërorët dhe i përfaqësojnë ato në fjali. Qëllimi i përdorimit të përemrit është të shmanget ripërdorimi i tyre. Përemri në përgjithësi ka karakteristikat gramatikore të gjinisë, numrit dhe rasës, por ka disa lloje përemrash të cilat nuk i kanë të tri këto kategoritë ose mund të jenë dhe të pandryshueshëm. Në gjuhën shqipe kemi shtatë lloje përemrash:

- *Përemri vetor* – unë, ti, ai, ajo, ne, ju, ata/ato.
  Është përemri që tregon veta të ndryshme. Ato kanë tri veta dhe dy numra. Vetat janë: veta e parë - që flet, veta e dytë - me të cilën ne flasim dhe veta e tretë -për të

cilën flasim. Për çdo vetë kemi dy numra njëjës dhe shumës. Përveç kategorisë së vetës dhe numrit, të gjithë përemrat kanë dhe kategorinë gramatikore të rasës dhe vetëm përemrat në vetën e tretë kanë dhe karakteristikat e gjinisë, përkatësisht ai dhe ata në mashkullore dhe ajo dhe ata në femërore. Përemrat e tjerë përdoren në të njëjtën formë në të dy gjinitë.

Këto përemra në rasën kallëzore dhe dhanore kanë trajta të shkurtra të thjeshta: mua (më), ty (të), atij/asaj (i) dhe atë (e), neve (na), juve (ju), atyre (u) dhe ata/ato (i). Këto trajta të shkurtra përdoren njëra pranë tjetrës duke formuar trajtat e shkurtra të bashkuara.

- *Përemri vetvetor* – vetja, vetvetja.
  Është përemri që tregon vetën që përfaqëson njëkohësisht atë që e kryen një veprim dhe atë mbi të cilin bie një veprim. Ky përemër zakonisht përdoret për të treguar persona dhe sende shumë rrallë. Përemri vetvetor nuk ka karakteristikën e gjinisë, por përdoret në të njëjtën formë në të dyja gjinitë. Ky përemër përdoret vetëm në numrin njëjës, pra nuk ka formë të shumësit. Ka karakteristikën gramatikore të rasës dhe lakohet njëlloj si emrat e gjinisë femërore në trajtë të shquar në numrin njëjës.
- *Përemri dëftor* – ai, ajo, ata, ato, ky, kjo, këta, këto, i/e/të tillë, i/e/të këtillë, i/e/të atillë, etj.
  Tregon frymorë, sende, apo tipare të frymorëve a sendeve që ndodhen larg apo afër folësit. Përemrat ky, kjo përdoren për të treguar persona a sende në afërsi të folësit, përemrat ai, ajo kjo përdoren për të treguar persona a sende në largësi të folësit, kurse përemrat i tillë, i atillë dhe i këtillë përdoren për të treguar tipare sendesh që ndodhen si në largësi dhe në afërsi të folësit. Këto përemra kanë kategorinë gramatikore të gjinisë, numrit dhe rasës.
- *Përemri pronor* – im, yt, tij, saj, ynë, juaj, tyre, etj.
  Është përemri që tregon se kujt i përket vete i përket një send i caktuar. Këto përemra tregojnë përkatësi ndaj një vete të caktuar. Përemrat pronorë karakterizohen nga raporti "pronë-pronar". Kategoritë gramatikore të përemrave pronorë janë kategoria e vetës, gjinisë, numrit dhe rasës. Përemrat pronorë kanë kategorinë e vetës njëlloj si përemri vetor por në ndryshim nga këto përemra ata kanë forma gjinie për të tri vetat. Gjinia a përemrit pronor varet nga gjinia e emrit që ai përcakton ose përfaqëson njëlloj si mbiemrat. Raporti "pronë-pronar" luan rol vetëm në vetën e tretë. Kjo do të thotë që në vetën e parë dhe të dytë, pavarësisht nga gjinisë së pronarit do të kemi forma të ndryshme vetëm në varësi të gjinisë së pronës. Për sa i përket numrit, pronari përshtate në numër me emrin që përcakton ose përfaqëson, dhe për dy vetat e para kemi dy forma për çdo numër kurse për vetën e tretë kemi katër forma për njëjësin dhe pesë forma për shumësin.

- *Përemri pyetës* – kush, cili, cila, çfarë, sa, ç, etj.
  Përdoret për të pyetur për një frymor ose send, për përkatësinë e një tipari, për sasinë, etj. Përemra të ndryshëm pyetës kanë kategori të ndryshme gramatikore, p.sh përemri kush dhe sa kanë vetëm kategorinë e rasës, përemri cili dhe cila kanë kategorinë e gjinisë, numrit dhe rasës, përemri ç dhe çfarë nuk kanë kategori gramatikore.
- *Përemri lidhor* – i cili, e cila, që, çka, etj.
  Përdoret për të lidhur një fjali të varur përcaktore me një gjymtyrë të një fjalie tjetër apo me një fjali tjetër. Në gjuhën shqipe përemrat lidhorë mund të jenë të caktuar, si që, cili dhe çka dhe të pacaktuar si kush, i cili, ç, çfarë. Përemrat që, çka, ç nuk kanë kategori gramatikore, dhe përdoren vetëm në këtë formë. Përemri cili ka kategoritë gramatikore si përemri pyetës cili, të gjinisë, rasës dhe numrit. Përemri kush ka kategoritë gramatikore të rasës si përemri pyetës kush.
- *Përemër i pacaktuar* – kush, dikush, një, njeri, ndonjë, askush, diçka, gjithë, etj.
  Përdoret për të treguar një frymor a send në mënyrë të pacaktuar. Këto përemra nga ana morfologjike ndahen në të pandryshueshëm dhe në të ndryshueshëm. Disa nga këto përemra mund të tregojnë frymorë a sende por disa mund të përjashtojnë ato.

4.1.5. Folja

Folja është pjesë e ndryshueshme e ligjëratës që emërton një veprim. Ajo ka kategoritë gramatikore të mënyrës, kohës, vetës, numrit dhe diatezës. Folja është bërthama kryesore e një fjalie.

Në gjuhën shqipe foljet klasifikohen sipas kuptimit dhe funksionit në:

- Folje ndihmëse
- Folje gjysmëndihmëse
- Folje kalimtare dhe jo kalimtare

Foljet mund të jenë formën e shtjelluar ose në format e pashtjelluara. Format e pashtjelluara të foljeve janë:

- Pjesorja: lexuar;
- Përcjellorja: duke lexuar;
- Paskajorja: për të lexuar;
- Mohorja: pa lexuar.

Zgjedhimi është tërësia e formave që merr folja sipas mënyrës, kohës, vetës dhe numrit. Foljet zgjedhohen në zgjedhimin vepror dhe zgjedhim jo vepror. Në gjuhën shqipe kemi gjashtë mënyra, tetë kohë të kategorizuara në tri kohë kryesore e shkuar, e tashme dhe e ardhme, tri veta: e parë e dytë dhe e tretë dhe dy numra: njëjës dhe shumës.

Mënyrat e foljeve dhe kohët përkatëse të çdo mënyre janë:

- Mënyra dëftore;
    - Koha e tashme;
    - Koha e pakryer;
    - Koha e kryer e thjeshtë;
    - Koha e kryer;
    - Koha më se e kryer;
    - Koha e kryer e tejshkuar;
    - Koha e ardhme;
    - Koha e ardhme e përparme;
    - Koha e ardhme e së shkuarës;
    - Koha e ardhme e përparme e së shkuarës.
- Mënyra habitore;
    - Koha e tashme;
    - Koha e pakryer;
    - Koha e kryer;
    - Koha më se e kryer;
- Mënyra lidhore;
    - Koha e tashme;
    - Koha e pakryer;
    - Koha e kryer;
    - Koha më se e kryer;
- Mënyra dëshirore;
    - Koha e tashme;
    - Koha e kryer;
- Mënyra kushtore;
    - Koha e tashme;
    - Koha e kryer;
- Mënyra urdhërore – kjo kohë ka formë vetëm për vetën e dytë njëjës dhe shumës.

4.1.6. Ndajfolja

Ndajfolja është pjesë e pandryshueshme e ligjëratës, e cila përcakton një tipar të një gjendje apo veprimi, rrethanën në të cilën realizohet gjendja apo veprimi, ose shkallën ose intensitetin e një veprimi apo cilësie. Ndajfoljet ndahen si:

- Ndajfolje përcaktore:
    - Të mënyrës – tregon se si kryhet veprimi i shprehur nga folja dhe luaj funksionin e rrethanorit te mënyrës, p.sh. mirë, keq, bukur, pastër, thjesht, etj;.
    - Të sasisë – tregon sasinë ose intensitetin e shkallës së veprimit dhe luan funksionin e rrethanorit të sasisë, p.sh. pak, shumë, fort, së tepërmi, kaq.

- Ndajfolje rrethanore:
    - Të kohës – tregon kohen kur kryhet apo sa zgjat një veprimi apo gjendje që është shprehur nga folja dhe luan funksionin e rrethanorit të kohës, p.sh. sot, dje, përnatë, aty për aty, mbrëmë;
    - Të vendit – tregon vendi ku kryhet veprimi i shprehur që është shprehur nga folja dhe luan funksionin e rrethanorit të vendit, p.sh. përpara, lart, poshtë, gjëkundi, diku, ku, nga;
    - Të shkakut – përdoren si mjet lidhës të një fjalie të varur nga një fjali tjetër ose për të pyetur për shkakun a qëllimin e një veprimi, p.sh. pse, përse, sepse.

Ndajfolja ka kategorinë e shkallës njëlloj si mbiemri.

4.1.7. Parafjala

Është pjesë e pandryshueshme e ligjëratës që shpreh marrëdhënie sinktaksore varësie ndërmjet fjalëve, si një emër, një përemër, një numërori apo një togfjalëshi. Parafjala mund të jetë ndajfoljore ose emërore. Sipas strukturës morfologjike ato ndahen në:

- Të thjeshta – para, pas, tutje, faqe, mes, majë, etj.;
- Të përngjitura – përpara, përbri, nëpër, ndërmjet, etj.;
- Lokucione – ballë për ballë, rreth e rrotull, për arsye, në lidhje me, etj..

Parafjalë të caktuara mund të përdoren në rasa të caktuara dhe një parafjë përdoret vetëm në një rasë të caktuar. Përjashtim bëjnë vetëm parafjalët ndaj dhe për që përdoren në dy rasa.

4.1.8. Lidhëza

Është pjesë e pandryshueshme e ligjëratës që përdoret për të lidhur dy gjymtyrë të një fjalie apo për të lidhur dy fjali. Sipas funksionit që ato kanë lidhëzat ndahen në lidhëza bashkërenditëse dhe lidhëza nënrenditëse. Lidhëza bashkërenditëse përdoret për të lidhur elementët që i përkasin të njëjtës klase, ose të njëjtit funksion. Lidhëza bashkërenditëse janë: dhe, e, edhe, apo, ose, prandaj, kurse, qoftë, etj. Lidhëza nënrenditëse përdoret për të lidhur dy elementë apo fjali të cilat kanë marrëdhënie varësie midis tyre. Lidhëza nënrenditëse janë: që, nëse, kur, si, pasi, gjersa, teksa, që kur, etj..

Sipas strukturës morfologjike ato ndahen në:

- Të thjeshta – ose, apo, as, dhe, në , o, por, etj.;
- Të përngjitura – sado, ngado, kudo, porsa, kurse, etj.;
- Lokucione – qoftë ... qoftë, në qoftë se, edhe pse, me qëllim që, etj..

4.1.9. Pjesëza

Është pjesë e pandryshueshme e ligjëratës që përdoret për të shprehur nuanca kuptimore apo emocione të një grupi emëror, foljor apo fjalie në tërësi. Nuk kanë kategori gramatikore duke qenë se janë pjesë së pandryshueshme.

Sipas strukturës morfologjike ato ndahen në:

- Të thjeshta – vallë, jo, ja, bash, etj.;
- Të përngjitura – posi, pale, kushedi, etj.;
- Lokucione – jo që jo, vetëm që, as që, vetëm e vetëm, po se po, etj.

4.1.10. Pasthirrma

Është pjesë e pandryshueshme e ligjëratës me vlerë thirrmore që përdoret për të shprehur dëshira, ndjenja, ndijime të folëset, por duke mos i emërtuar ato. Nuk kanë kategori gramatikore duke qenë se janë pjesë së pandryshueshme. Shembuj pasthirrme janë: oburra, ua, bobo, he ... he, etj.

4.2. Sintaksa e gjuhës shqipe

Sintaksa është pjesa e gramatikës e cila studion mënyrat e bashkimit të fjalëve sipas kuptimit të tyre gramatikor në ligjërim, marrëdhëniet që vendosen midis njësive të fjalisë dhe rolin që fjalët kanë në fjali. Fjalia është një grup i lidhur fjalës sipas kuptimit që fjalët kanë dhe rregullave gramatikore të gjuhës. Në gjuhën shqipe dy janë grupet kryesore përbërëse të fjalisë, grupi emëror që tregon personat për të cilat flitet dhe grupi foljor që tregon çfarë flitet për personat e treguar nga grupi emëror. Në gjuhën shqipe kemi katër tipe fjalish që i kemi trajtuar në çështjen 4.2.1 dhe gjashtë pjesë përbërëse të fjalisë që i kemi trajtuar në çështjen 4.2.2.

4.2.1. Tipet e fjalive

Fjalia është një grup fjalësh të lidhura bazuar në kuptimin dhe rregullat gramatikore të gjuhës shqipe. Fjalitë në gjuhën e folur përcaktohet nëpërmjet një intonacioni të veçantë që mund të jetë: ngjitës, zbritës, etj.. Kurse në gjuhën e shkruar, fjalia përcaktohet nga një shenjë pikësimi që përcakton fundin e saj që mund të jetë: pikë, pikëçuditje, pikëpyetje, pikëpresje. Fjalia mund të jetë e thjeshtë kur ka vetëm një folje ose e përbërë kur ka më shumë se një folje. Fjalitë mund të jenë në formën: thirrmore, veprore ose pësore, dhe pohore ose mohore. Nëse folja e fjalisë është në formën veprore dhe fjalia është në formën veprore. Nëse folja e fjalisë është në formën joveprore, fjalia është në formën pësore. Fjalia në formën pohore pëson një shndërrim mohor kurse në formën mohore, mohon një pohim. Fjalia është thirrmore kur paraqet një ndjenjë habie, kënaqësie, frike, zemërimi, etj.. Në gjuhën shqipe kemi tre tipe fjalish si në vijim.

Fjalia dëftore

Fjalia dëftore tregon fakte të realitetit, jep informacion rreth një vëzhgimi, fakti, gjykimi apo opinioni. Këto janë fjalitë më të përdorshme. Në rastin kur është fjali e mëvetësishme, ajo shqiptohet me intonacion tregues. Në fund të fjalisë dëftore toni ulet dhe është i lartë në fjalën që mban theksin logjik. Zakonisht folja në fjalinë dëftore është në mënyrën dëftore, por ka raste kur përdoret në këto fjali dhe folje në mënyrën habitore, lidhore apo kushtore.

Fjalia dëshirore dhe nxitëse

Fjalia dëshirore dhe nxitëse tregojnë vullnetin e folësit në formë dëshire apo kërkese për të realizuar ose jo një veprim të caktuar. Fjalia dëshirore shpreh dëshirën e folësit për të realizuar apo jo një veprim kurse fjalia nxitëse shpreh nxitjen, ndikimin e folësit në realizimin ajo të një veprimi. Fjalia nxitëse nxit për zbatimin e një këshille, urdhri apo kërkese. Në fjalinë dëshirore folja mund të jetë në mënyrën dëshirore ose lidhore. Në fjalinë nxitëse folja mund të jetë në mënyrën dëftore, lidhore ose urdhërore.

Fjalia pyetëse

Janë fjali me anë të së cilave folësi pyet dhe do të mësojë të cilën ai nuk e di saktësisht apo nuk di vërtetësinë e saj. Fjalitë pyetëse përdoren gjatë bashkëbisedimit, ku një nga folësit pyet dhe pret përgjigje nga një apo disa folës të tjerë. Ato mund të jenë:

- fjali pyetëse tërësore – folësi i përgjigjet pyetjes me po, jo ose foljen që është përdorur në fjalinë pyetëse;
  P.sh. Ishe të shtunën në kinema? Po/jo/isha/nuk isha.
- fjali pyetëse të pjesshme – folësi ka informacion rreth asaj që pyet por do të dijë më shumë informacion rreth një fakti që nuk e di;
  P.sh. Ç'ditë u mbajt konferenca? – Të premten.
- fjali pyetëse retorike – folësi bën pyetje por nuk pret të marrë një përgjigje.
  P.sh. Ku i del dot atij ti?

4.2.2. Gjymtyrët e fjalisë

Gjymtyrët kryesore të fjalisë janë kryefjala, kallëzuesi dhe pjesët plotësuese: kundrinori, rrethanori dhe përcaktori.

Kryefjala

Kryefjala është gjymtyrë kryesore emërore ose përemërore e fjalisë dhe ka marrëdhënie kallëzuesore me kallëzuesin. Për të gjetur kryefjalën të një fjalie përdoren pyetjet: kush?, cili?, cila? dhe cilët?. Kryefjala mund të tregoj një ose disa frymorë ose sende. Zakonisht kryefjala qëndron para foljes. Kryefjala mund të shprehet me:

- emër në rasën emërore, në formë të shquar ose të pashquar, në numrin njëjës ose shumës;

- emër në rasën gjinore;
- grup emëror;
- përemër dëftor, lidhor, të pacaktuar, vetor, pyetës në çdo vetë dhe numër;
- numëror;
- pjesë të tjera të ligjëratës të emërzuara;
- togfjalësh;

Kallëzuesi

Kallëzuesi është gjymtyrë kryesore foljore e fjalisë dhe ka marrëdhënie kallëzuesore me kryefjalën. Ai shpreh një gjendje apo veprim të kryefjalës. Për të gjetur kallëzuesin përdoren pyetjet: ç' është? dhe ç'bën?. Kallëzuesi përshtate në vetë, numër me kryefjalën dhe qëndron zakonisht pas saj. Kallëzuesi është foljor dhe emëror. Kallëzuesi foljor mund të jetë i thjeshtë foljor i shprehur nëpërmjet një forme foljore të çdo diateze ose i përbërë foljor i shprehur nëpërmjet dy fjalëve kuptimplota, ku fjala e parë tregon mënyrën e veprimit dhe fjala e dytë tregon një veprim. Kallëzuesi emëror përbëhet nga gjymtyrë pjesë: këpuja dhe gjymtyra emërore. Si këpujë përdoret folja jam kurse si gjymtyrë emërore përdoret një emër, një tog fjalës, një përemër, një mbiemër apo një paskajore që nga ana leksikore shprehin një kallëzues emëror.

Kundrinori i drejtë

Kundrinori është gjymtyrë e dytë e fjalisë dhe tregon objektin, që mund të jetë frymor ose jo, mbi të cilin bie veprimi i shprehur nga folja. Ai vihet në atë rasë që kërkon gjymtyra prej së cilës varet. Për të gjetur kundrinorin e drejtë përdoren pyetjet kë?, cilin?, cilën? Dhe çfarë?. Kundrinori i drejtë shprehet nëpërmjet një grupi emëror pa parafjalë në rasën kallëzore *dhe* lidhet me folje në formën veprore. Nëse një fjali veprore kthehet në një fjali pësore atëherë fjalia do të pësojë këto ndryshime:

- Kryefjala do të bëhet kundrinor i drejtë;
- Kundrinori i drejtë do të bëhet kryefjalë.

Kundrinori i drejtë shprehet me:

- emër zakonisht në formë të shquar në rasën kallëzore;
- grup emëror;
- përemër;
- trajtë të shkurtër të përemrit vetor;
- numëror;
- pjesë të tjera të emërzuara të ligjëratës;
- togfjalësh;

Kundrinori i zhdrejtë

Kundrinori i zhdrejtë ka përdorim të gjerë e larmishmëri në përdorim si nga ana kuptimore dhe nga format gramatikore. Ky kundrinor ndërtohet me folje kalimtare dhe jo kalimtare, në veprore dhe joveprore. Llojet e kundrinorit të zhdrejtë janë:

- Kundrinori i zhdrejtë pa parafjalë;
- Kundrinori i zhdrejtë me parafjalë.

Kundrinori i zhdrejtë pa parafjalë

Kundrinori i zhdrejtë pa parafjalë është gjymtyre e dytë e fjalisë e cila lidhet në mënyrë të drejtpërdrejtë me foljen. Kundrinori i zhdrejtë pa parafjalë tregon frymorin apo sendin mbi të cilin bie veprimi i shprehur nga folja. Ky kundrinor shprehet me ato pjesë të ligjëratës që shprehet dhe kundrinori i drejtë por në trajtën e rasës dhanore. Në shumicën e rasteve ky kundrinor shoqërohet nga trajta e shkurtër e përemrit vetor të rasës dhanore. Për të gjetur kundrinorin e zhdrejtë pa parafjalë përdoren pyetjet: kujt?, cilit?, cilës?, cilëve?, cilave?. Zakonisht ky kundrinor në fjali vendoset pas foljes por mund të ndryshojë pozicion lirshmërisht. Kundrinori i zhdrejtë pa parafjalë shprehet me:

- emër në rasën dhanore;
- grup emëror;
- përemër.

Kundrinori i zhdrejtë me parafjalë

Kundrinori i zhdrejtë me parafjalë, përdoret gjithmonë me një parafjalë dhe ka përdorim më të gjerë se kundrinori i zhdrejtë pa parafjalë. Ky kundrinor nuk është i lidhur me foljen aq sa është kundrinori i drejtë. Ai nuk është pjesë kryesore e fjalisë edhe pse plotëson më mirë kuptimin e foljes. Në një fjali folja mund të qëndroj dhe pa kundrinorin e zhdrejtë me parafjalë. Kundrinori i zhdrejtë me parafjalë varet jo vetëm nga folja por edhe nga pjesë të tjera të ligjëratës. Ky kundrinor plotëson jo vetëm folje kalimtare në veprore, mesore apo pësore, por dhe folje jokalimtare, një përemër, një mbiemër, një numëror, një togfjalësh apo dhe pjesë të tjera të ligjëratës të emërzuara. Për të gjetur kundrinorin e zhdrejtë me parafjalë përdoren pyetjet: për kë?, me se?, nga kush?, prej kujt?,nga se?. Parafjalët më të zakonshme në ndërtimin e kundrinorit të zhdrejtë me parafjalë janë: te, me, nga, mbi, pa, kundër, nën, etj.. Kundrinori i zhdrejtë me parafjalë mund të ndërtohet dhe me anë të shprehjeve parafjalore, si: me anë të, bashkë me, në lidhje me, etj.. Kundrinori i zhdrejtë me parafjalë shprehet me:

- emër;
- grup emëror i shprehur me një emër, një përemër ose me një emër dhe një përcaktor në rasën emërore, kallëzore ose rrjedhore;
- përemër .

Rrethanori

Rrethanori është gjymtyrë e dytë e fjalisë i cili ka një marrëdhënie rrethanore me gjymtyrën kryesore të fjalisë prej të cilës varet dhe tregon rrethanat e kryerjes së veprimit. Rrethanori tregon: kohë, vend, shkak, mënyrë, qëllim dhe sasi. Për të gjetur rrethanorin përdoren pyetjet: kur?, ku?, si?, pse?, sa?, përse?. Pozicioni në fjali i rrethanorit është pas folje ose kallëzuesit por shpesh mund të dalë dhe para tyre.

Rrethanori i kohës

Ky rrethanor tregon kohën e realizimit të një veprimi. Për të gjetur rrethanorin e kohës përdoret pyetja: kur?. Ai mund të lëviz lirshëm në pozicione të ndryshme në fjali. Ky rrethanor shprehet me:

- emër me e pa parafjalë;
- grup emëror;
- ndajfolje ose shprehje ndajfoljore;
- formë të pashtjelluar;

Rrethanori i vendit

Ky rrethanor tregon vendin e realizimit të një veprimi. Për të gjetur rrethanorin e vendit përdoret pyetja: ku?. Ai mund të lëviz lirshëm në pozicione të ndryshme në fjali. Ky rrethanor shprehet me:

- emër me parafjalë: p.sh. Librat i mbaj në bibliotekë. (ku?)
- grup emëror: p.sh, Jam nisur për në Galerinë e Arteve. (për ku?)
- ndajfolje: p.sh, Zhurma dëgjohej sipër. (ku?)

Rrethanori i shkakut

Ky rrethanor tregon shkakun e një veprimi të caktuar. Rrethanori i shkakut varet nga kallëzuesi por mund të varet dhe nga një emër ose mbiemër. Ai vendoset pas gjymtyrës nga e cila varet. Për të gjetur rrethanorin e shkakut përdoret pyetja: pse?. Ky rrethanor shprehet me:

- emër me parafjalë;
- grup emëror;
- ndajfolje;
- përcjellore.

Rrethanori i mënyrës

Ky rrethanor tregon mënyrën, cilësinë dhe shkallën e veprimit. Për të gjetur rrethanorin e mënyrës përdoret pyetja: si?. Ai mund të lëviz lirshëm në pozicione të ndryshme në fjali. Ky rrethanor shprehet me:

- grup emëror;
- ndajfolje ose shprehje ndajfoljore;

- formë të pashtjelluar.

Rrethanori i qëllimit

Ky rrethanor tregon qëllimin e kryerjes së veprimit dhe varet nga kallëzuesi. Rrethanori i qëllimit vendose pas gjymtyrës nga e cila varet. Për të gjetur rrethanorin e qëllimit përdoret pyetja: përse?. Ky rrethanor shprehet me:

- emër me parafjalë;
- grup emëror;
- ndajfolje;
- paskajore.

Rrethanori i sasisë

Ky rrethanor tregon sasi, karakteristika sasiore, masën e kohës, gjendjes, hapësirës dhe shkallën e mënyrës apo cilësisë së një veprimi. Për të gjetur rrethanorin e sasisë përdoret pyetja: sa?. Ky rrethanor shprehet me:

- grup emëror;
- ndajfolje ose shprehje ndajfoljore.

Përcaktori

Përcaktori është gjymtyrë e dytë e fjalisë dhe ka marrëdhënie përcaktore me gjymtyrën e fjalisë prej të cilës varet. Ai përcakton, saktëson apo sqaron gjymtyrën me të cilën varet. Gjithmonë në funksionin e përcaktorit është emri në rasën gjinore të shquar. Përcaktori shprehet me: emër, përemër dhe mbiemër. Në gjuhën shqipe bazuar në llojin e lidhjes sintaksore kemi katër lloje përcaktorësh:

- përcaktori me përshtatje, që përcakton cilësinë e emrit bërthamë;
- përcaktori me drejtim, që tregon një tipar të emrit që ai përcakton;
- përcaktori me bashkim;
- përcaktori me ndajshtim, që emërton emrin bërthamë.

4.3. Vështirësitë e etiketimit në gjuhën shqipe

Gramatika e gjuhës shqipe është mjaft komplekse dhe e pasur me forma të fjalës. Këto janë faktorët që e bëjnë etiketimin e tekstit në gjuhën shqipe një sfidë. Më poshtë kemi listuar disa nga faktorët që ndikojnë në vështirësinë e etiketimit:

- **Dykuptimshmëria:** Një numër shumë i madh fjalës në gjuhën shqipe mund t'i përkasin pjesëve të ndryshme të ligjëratës dhe dallohen vetëm në varësi të kontekstit që janë përdorur. Për shembull:
  Sa persona janë? → Sa është përemër pyetës .
  Sa kishte mbërritur. → Sa është ndajfolje kohe.
  E donte **sa** edhe të motrën. → Sa është lidhëz nënrenditëse.

- **Rend i papërcaktuar në fjali:** Në gjuhën shqipe fjalët në fjali nuk kanë një rendë të mirë përcaktuar varësi të funksionit sintaksor që ato kanë.
- **Forma të shumëllojta:** Në gjuhën shqipe fjalët gjatë lakimit pësojnë shumë ndryshime. Për shembull një folje mund të ketë shumë forma leksikore në varësi të karakteristikave morfologjike gjatë lakimit.
- **Nuk ka rregulla në lakim:** Gjatë lakimit fjalë të ndryshme të cilat i përkasin të njëjtës pjesë së ligjëratës pësojnë ndryshime të cilat jo domosdoshmërisht ndjekin gjithmonë një rregull të caktuar.
- Një problem në tekstet shqip në mediet online është mospërdorimi i germës ë dhe ç të cilat përdoren si germa e dhe c.

KREU 5

ETIKETIMI I PJESËVE TË LIGJËRATËS DHE GJETJA E RRËNJËS E TEMËS SË FJALËS

Në këtë kapitull trajtohet në detaje sistemet që përdoren për etiketimin morfologjik dhe për gjetjen e temës dhe rrënjës së fjalës. Këto mjete kanë një rëndësi të veçantë në shumë fusha dhe në shumë detyra të përpunimit të gjuhës natyrale dhe analizimit të tekstit. Më tej, në këtë kapitull diskutohet me detaje skema Universal Dependencie (UD) dhe parseri Turku Neural Parser Pipeline, që janë përdorur në këtë disertacion për krijimin e një korpusi të etiketuar në gjuhën shqipe dhe realizimin e një etiketuesi morfologjik dhe temëzuesi për gjuhën shqipe.

5.1. Hyrje

Në fushën e përpunimit të gjuhës natyrale, sistemet e etiketimit morfologjik apo gjetjes së rrënjës dhe temës kanë një rëndësi të veçantë dhe përdoren gjerësisht në aplikime dhe në fusha të ndryshme. Këto sisteme zakonisht përbëjnë hapin e parë të përpunimit apo parapërpunimit të tekstit për t'u përdorur në aplikime të tjera si, nxjerrje informacioni, klasifikim teksti, sistemet e pyetje përgjigjeve, analizimin e ndjenjave, gjetjen e fjalëve kyçe, etj. Në këto sisteme është e rëndësishme të dihet se çfarë përfaqëson nga ana gramatikore një fjalë apo të përcaktohet forma bazë e fjalëve të lakuara.

Përcaktimi i një etikete të çdo fjale të një sekuence fjalësh në hyrje, ku sekuenca e hyrjes dhe ajo e daljes kanë të njëjtën gjatësi, të njëjtin numër fjalësh, përcaktohet si etiketimi i sekuencës (angl. sequence labeling) (Jurafsky & Martin, 2020).

Sistemet e etiketimit morfologjik, etiketuesi i pjesëve të ligjëratës (angl. Part-of-Speech Tagger, (POS)) dhe etiketuesi i karakteristikave morfologjike, kanë për qëllim të përcaktojnë për çdo fjalë në tekst një etiketë që përkon me kategorinë e pjesës së ligjëratës që i përket fjala dhe etiketat që përkojnë me karakteristikat morfologjike të formës së fjalës. Etiketimi bazohet si në formën aktuale të fjalës edhe në kontekstin që ajo është përdorur, në lidhje me fjalët e tjera në shprehjen apo dhe në fjalinë në të cilën ajo ndodhet. Në procesin e etiketimit hyrja është një sekuencë fjalësh X_1, X_2, ..., X_n dhe një grup etiketash dhe dalja është një sekuencë etiketash Y_1, Y_2, Y_3, ..., Y_n, ku çdo etiketë Y_i i korrespondon përkatësisht vetëm një fjale X_i të sekuencës së hyrjes (Jurafsky & Martin, 2020).

Nëse në hyrje të etiketuesit kemi fjalinë:

Si thua?

Në dalje do të kemi fjalinë ku është etiketuar çdo fjalë e saj:

Si ADV AdvType=Man
thua VERB Mood=Ind|Number=Sing|Person=2|Tense=Pres|Voice=Act
? PUNCT _ _ _ _ _ _

Parseri ose etiketuesi sintaksor ka si qëllim të ndërtojë pemën e varësisë së fjalëve në një fjali. Në këto sisteme një fjali paraqitet në formën e një peme duke identifikuar lidhjet/varësitë që kanë midis tyre fjalët.

Sistemet e gjetjes së rrënjën së fjalës, (angl. stemmer), kanë për qëllim të transformojnë një fjalë nga forma aktuale në rrënjën e tyre, pra identifikimi i rrënjës së një fjale. Në anglisht procesi i gjetjes së rrënjës së fjalës përcaktohet si stemming, në shqip do t'i referohemi si rrënjëzim. Në anglisht sistemi i gjetjes të rrënjës së fjalës quhet stemmer dhe në shqip do t'i referohemi si rrënjëzues.

Sistemet e gjetjes së temës së fjalës, (angl. lemmatizer) kanë si qëllim të transformojnë një fjalë nga forma aktuale në temën e saj që është fjala që gjendet në fjalor, pra identifikimi i temës së një fjale. Në anglisht procesi i gjetjes së temës së fjalës përcaktohet si lemmatizing dhe në shqip do t'i referohemi si temëzim. Në anglisht sistemi i gjetjes të rrënjës së fjalës quhet lemmatizer dhe në shqip do t'i referohemi si temëzues. Temëzimi dhe rrënjëzimi është një proces i rëndësishëm për motorët e kërkimit apo për studime të ndryshme gjuhësore sidomos për gjuhë të pasura me forma gramatikore të fjalëve.

Nëse në hyrje të temëzuesit kemi fjalinë:

Si thua?

Në dalje do të kemi temat e çdo fjale të fjalisë:

Si them ?

Një nga problemet kryesore në sistemet e etiketimit morfologjik dhe parsimit është dykuptimësia. Në fjali të ndryshme e njëjta fjalë mund të përdoret si pjesë e ndryshme e ligjëratës, p.sh. mund të paraqitet si emër ose si mbiemër.

*1) Ky det është i **kaltër**.* *2) Ngjyej me të **kaltër.***

Në dy fjalitë e mësipërme kemi përdorimin e të njëjtës fjalë "kaltër", por në fjalinë e parë fjala "kaltër" është një mbiemër kurse në fjalinë e dytë kjo fjalë është një emër. Pikërisht një etiketues i mirë duhet të etiketojë saktë fjalën në varësi të funksionit që ajo ka në fjali dhe të dallojë etiketimin e fjalëve të cilat mund të përdoren në të njëjtën formë por të kenë funksion gramatikor të ndryshëm.

*1) Dëgjoj më mirë nga **veshi** i djathtë.* *2) Ajo **veshi** një fustan me pika.*

Në dy fjalitë e mësipërme kemi përdorimin e të njëjtës fjalë "veshi", por në fjalinë e parë fjala "veshi" është emër kurse në fjalinë e dytë kjo fjalë është folje.

Në sistemet e rrënjëzimit apo temëzimit një nga problemet është gjetja saktë e rrënjës apo temës së fjalëve që mund të kenë të njëjtën formë por rrënjë ose temë të ndryshme.

Nëse kemi dy fjalitë:

1) *Ju **qetë** në Milano shumë herë.* *2) Lagjja ishte shumë e **qetë**.*

Në fjalinë e parë fjala qetë është folje dhe tema e saj është jam, kurse në fjalinë e dyte kjo fjalë është mbiemër dhe tema e saj është qetë.

Një nga parametrat më të rëndësishëm për vlerësimin e performancës së këtyre sistemeve është saktësia (angl. accuracy) e etiketimit, gjetjes së rrënjës apo temës.

Saktësia përcakton se sa herë sistemi ka realizuar parashikim të saktë të etiketës, rrënjës apo temës së fjalës sipas etiketimit të realizuar nga ekspertët e fushës. Llogaritet si raport i numrit të etiketave, rrënjëve apo temave të përcaktuara saktë nga sistemi me numrin e të gjithë etiketave, rrënjëve apo temave të përcaktuara. Formula e llogaritjes është:

$$Saktësia = \frac{nr\_etiketave\_përcaktuar\_saktësisht}{nr\_total\_të\_etiketave}$$

Etiketuesit e pjesëve të ligjëratës më të mirë për gjuhën angleze raportohen të kenë saktësi etiketimi rreth 97% që është dhe saktësia e ekspertëve të fushës (Jurafsky & Martin, 2020).

Gjithashtu performanca e këtyre sistemeve mund të vlerësohet në terma të shpejtësisë së etiketimit, parsimit, rrënjëzimit apo temëzimit apo sasisë së memories së shfrytëzuar gjatë përpunimit, etj.

Teknikat e përdorura për implementimin e mjeteve mund të jenë të bazuara në të mësuarin e kontrolluar ose e pakontrolluar ose hibride. Në dy çështjet në vazhdim do të trajtohen më me detaje teknikat e përdorura për etiketim morfologjik, parsim, rrënjëzim dhe temëzim.

5.2. Teknikat e etiketimit morfologjik e parsimit

Zhvillimi i etiketuesve morfologjik automatik dhe parsuesve i ka fillesat që në fund të viteve 1950, por vetëm 20 vitet e fundit do të bëheshin zhvillimet më të mëdha. Etiketuesit dhe parsuesit e parë automatik kanë qenë të bazuar në rregulla gramatikore të implementuara nëpërmjet shprehjeve të rregullta në automata të fundme apo si një set rregullash të renditura sipas kontekstit, ato që në anglisht u referohemi si rule-based taggers.

Si një ndër eriketuesit më të hershëm të pjesëve të ligjëratës për gjuhën angleze mund të konsiderohet etikuesi, pjesë e parserit të zhvilluar nga Harris (1962) në kuadër të projektit Transformations and Discourse Analysis Project (TDAP) (Jurafsky & Martin, 2020). Parseri është implementuar si një kaskadë *finite-state transducers* (fst). Parseri konsiston në 7 faza ku në fazën e parë çdo fjale i përcaktohet një ose më shumë etiketa POS të cilat renditen sipas frekuencës së tyre mbi bazën e një përcaktimi në fjalor, në fazën e dytë idiomat zëvendësohen më një etiketë të vetme POS dhe në fazën e tretë përdoren 14 rregulla të shkruara për të menaxhuar dykuptimshmërinë në etiketimin e pjesëve të ligjëratës. Fazat nga e katërta në të gjashtë janë fst të cilat përdoren për përpunimin e

frazave të tipeve të ndryshme duke gjeneruar në fazën e fundit një stringë që përcakton marrëdhënien e duhur midis fjalëve (Joshi & Hopely, 1996).

Dy sisteme etiketimi të zhvilluar më vonë por që përdorin të njëjtën filozofi janë CGC dhe TAGGIT. Të dy këto sisteme bazohen në një arkitekturë me dy faza për të përcaktuar etiketën e pjesëve të ligjëratës. Në fazën e parë përdoret një fjalor për t'i përcaktuar një fjale një set etiketash të pjesëve të ligjëratës të mundshme. Në fazën e dytë implementohen një set rregullash të shkruara për të menaxhuar dykuptimshmërinë dhe për t'i përcaktuar etiketën e duhur fjalës. CGC përdor një fjalor prej 1500 fjalësh dhe rreth 500 rregulla kurse TAGGIT përdor një fjalor më të zgjeruar me 87 etiketa dhe rreth 3300 rregulla për zgjidhjen e problemit të dykuptimshmërisë. Përdorimi në këto sisteme të një bashkësi rregullash për të shmangur dykuptimshmërinë për një gjuhë të caktuar bën të pamundur përdorimin e këtyre sisteme në gjuhë të tjera. Në mënyrë që këto sisteme të përdoren për një gjuhë tjetër nga e cila janë implementuar fillimisht do të duhet të rishkruhen rregullat sipas gjuhës në të cilën do përdoren (Jurafsky & Martin, 2020).

Më tej në vitet 1970 kemi aplikimin e metodave stokastike në zhvillimin e këtyre sistemeve. Aplikimi i këtyre metodave solli zhvillimin e sistemeve më perfomante. Modelet e hapura Markoviane (MM), Modelet e Fshehura Markoviane (angl. Hidden Markov Models (HMM)), algoritmi i Viterbit dhe Maximum Entropy (ME) janë metodat më të përdorura në etiketuesit stokastike. Aplikimi këtyre metodave i bëri sistemet të pavarura nga gjuha por të varura nga një korpus i etiketuar. Janë propozuar zgjidhje të ndryshme, ku mund të përmendim etiketuesin CLAWS i cili mund të konsiderohet si një implementim i thjeshtë i algoritmit HMM dhe etiketuesin PARTS që mund të konsiderohet si implementimi i parë i një etiketuesi të bazuar në HMM (Jurafsky & Martin, 2020).

Etiketuesi TnT (Brants, 2000) është një etiketues probabilitar për etiketimin e pjesëve të ligjëratës. Ai është një implementim i algoritmit Viterbi për një model të rendit të dytë HMM. Algoritmi Viterbi dhe modeli HMM përdorin modele skolastike dhe kur përdoren për etiketim të pjesëve të ligjëratës, ka si qëllim të përcaktojë etiketën më të mundshme, pra me probabilitet më të lartë, për një fjalë. Duke u bazuar në dy modele linguistike, atë leksikor dhe kontekstual, ky model ka performancë të mirë në etiketim. Modeli mëson nga një korpus i etiketuar për të gjeneruar etiketën më të mundshme të një fjale të paetiketuar. Autorët kanë propozuar dy modele të bazuara në këtë etiketues, një për gjuhën angleze me saktësi 94.5% dhe një në gjuhën gjermane me saktësi më të lartë prej 96.7%. Më tej është implementuar dhe një model duke përdorur një nga korpuset e etiketuara më të përdorshëm për gjuhën angleze Penn Treebank me saktësi prej 96.7%. Ky korpus është përdorur për të ndërtuar model nga etiketues të tjerë probabilitar apo të bazuar në MA, rrjete neurale apo MTh. Etiketuesi TnT raportohet të ketë saktësinë më të ulët në këtë korpus.

Stanford Log-linear Part-Of-Speech Tagger (Toutanova, et al., 2003) është një etiketues i bazuar në algoritmin maximum entropy (ME) që nuk është përdorur më parë për implementimin e sistemeve të tilla. Etiketuesi mëson sipas një modeli të kushtëzuar log-linear nga të dhëna të etiketuara duke përdorur metodën maximum entropy. Ideja e

modelimit ME është të zgjidhet probabiliteti i shpërndarjeve p që ka entropinë më të lartë në bashkësinë e shpërndarjeve që kënaqin disa kufizime të përcaktuara. Kufizimet e detyrojnë modelin që të sillet sipas karakteristikave të mësuara nga të dhënat e trajnimit. Autorët kanë aplikuar edhe karakteristika më të sofistikuara linguistike dhe karakteristika që modelojnë edhe parashikimet e realizuara më parë për të marrë një model me performancë më të lartë. Korpusi Penn Treebank është përdorur për të krijuar një model etiketuesi të bazuar në Stanford Log-linear POS për anglishten me një saktësi prej 97.24%. Krahasuar me modelin e TnT të trajnuar mbi të njëjtin korpus, Stanford Log-linear POS ka performancë më të lartë.

Një ndër etiketuesit më të thjeshtë dhe më të përdorur për gjuhën angleze është etiketuesi i propozuar nga Brill (1992). Etiketuesi i propozuar bazohet në një metodë të thjeshtë etiketimi bazuar në transformim (angl. Transformation-Based POS tagging) që është një kombinim i teknikave të bazuara në rregulla dhe teknikave stokastike (Brill, 1995). Skema e etiketimit e përdorur në këtë etiketues është Pen Treebank, ku janë përcaktuar 36 etiketa. Procesi i etiketimit është bazuar në rregulla të cilat janë të mësuara nga një korpus i etiketuar dhe kalon në tri faza. Në fazën e parë etiketuesi i përcakton një fjale etiketën më të mundshme duke u bazuar në të mësuarin nga korpusi i etiketuar. Në këtë fazë nëse një fjalë nuk është hasur më parë, do t'i përcaktohet një etiketë duke u bazuar në etiketën më të përdorshme në fjalët që kanë përputhje me tre karakteret e fundit të fjalës që po etiketohet. Në fazën e dytë algoritmi përcakton etiketën më të mundshme për të qenë etiketa e saktë duke kontrolluar nëse mund të aplikohet çdo transformim i mundshëm që është identifikuar gjatë fazës së trajnimit. Dhe në fazën e fundit ripërsëriten dy fazat e para derisa plotësohet një kusht i caktuar. Ekzistenca e këtij cikli përsëritje mund ta çojë algoritmin në një cikël të pafundmë gjatë ekzekutimit. Gjatë trajnimit të modelit për etiketim 90% e korpusit të etiketuar përdoret për trajnim, 5% për gjenerimin e rregullave të transformimit dhe 5 % për testim. Modeli i këtij etiketuesi kërkon sasi të vogla memorie për t'u ekzekutuar dhe identifikimi i gabimeve është shumë i thjeshtë. Procesi i etiketimit me anë të këtij etiketuesi është shumë i avashtë dhe kjo përbën një problem.

Në 20 vitet e fundit teknikat e përdorura në implementimin e etiketuesve morfologjikë dhe parsuesve janë të shumëllojshme por një fokus të veçantë në fushën akademike i është dhënë zhvillimit të modeleve të bazuara në rrjete neurale dhe të mësuarit e thelluar (Voutilainen, 2005). Një aplikim të gjerë kanë patur dhe algoritme të ndryshme MA si, Decision Tree, Bayesian Nets, Case-Based, Inductive Logic dhe Suport Vector Machines (Tufiș & Ion, 2014). Në shumicën e implementimeve të realizuara janë sisteme të cilat bazohen në të mësuarit e automatizuar të kontrolluar dhe kanë nevojë për një korpus të etiketuar për të mësuar. Këto janë dhe sistemet më përformante aktualisht.

Modelet e etiketuesve dhe parsuesve të bazuara në rrjete neurale mund t'i klasifikojmë në dy tipe: modele të bazuara në tranzicion (angl. transition-based models) dhe modele të bazuara në graf (angl. graph-based models). Për të ndërtuar pemën e parsimit, një parsues i bazuar në një model tranzicioni e përpunon fjalinë fjalë-pas-fjale nga e majta në të djathtë.

Në këtë rast pema e parsimit ndërtohet hap-pas-hapi. Parseri mëson të parashikojë lidhjen nga një hap në tjetrin nëpërmjet një historiku të parsimeve të mëparshme. Përcaktimi i pemës përfundimtare të parsimit realizohet duke ruajtur lidhjet e parashikuara të çdo hapi derisa të jetë përcaktuar pema komplete e varësisë. Koha e nevojshme për ndërtimin e pemës së varësisë është lineare në lidhje me gjatësinë e fjalisë që po etiketohet. Duke qenë se vendimmarrja e çdo hapi bazohet në një pemë parsimi të ndërmjetme, është e nevojshme përdorimi i metodave speciale të trajnimit të cilat janë të avashta. Kurse modeli i bazuar në graf mëson në një hap të vetëm dhe më pas realizon një kërkim shterues në të gjithë hapësirën e pemës së varësisë për të gjetur atë më përfaqësuesen të fjalisë. Në këtë metodë modeli mëson parametrat e duhur për të përcaktuar grafet e ndërmjetëm të saktë duke faktorizuar këto grafe mbi bazën e lidhjeve të harqeve të veçantë dhe analizimin e tyre për të përcaktuar grafin me peshë më të madhe për fjalinë. Kjo mënyrë funksionimi sjell përmirësime të performancës së parsuesit, veçanërisht në rastet e fjalive komplekse por është një metodë shumë e avashtë (Li, et al., 2020; McDonald & Nivre, 2011). Studime të ndryshme përcaktojnë që edhe pse të dy modelet janë shumë të ndrysëm nga ana teorike nuk kanë ndonjë ndryshim të dukshëm në saktësi për gjuhë të ndryshme (McDonald & Nivre, 2011). Qasje të ndryshme janë propozuar për të përmirësuar problemet që shfaqin të dy modelet. Këtu mund të theksojmë: grupimin – duke peshuar parashikimet e sistemeve të ndryshme, integrimi i karakteristikave - duke kombinuar dy modelet, ku dalja e një modeli përdoret për të përcaktuar karakteristikat që do të përdoren në modelin tjetër dhe qasje të cilat modifikojnë strukturën e modelit duke realizuar parsera të bazuar në tranzicion të trajnuar globalisht ose parsera të bazuar në graf me karakteristika të reja (Zheng, 2017).

Li et al. (2020) kanë propozuar një parser me tre komponentë, një enkoder për përpunimin e fjalive në hyrje, një shënjues i modifikuar që ka si qëllim përcaktimin e rendit të parsimit dhe një modul *greedy* që përcakton pemën e varësisë. Ky parser kombinon *greedy parsing inference* me *global arc scoring* të modelit të bazuar në graf në vend të *local feature scoring* të modelit të bazuar në tranzicion. Parseri është i shpejt në trajnim dhe dekodim dhe ka saktësi të lartë.

Në implementimin e parserit INDP janë marrë në mënyrë eksplicite në konsideratë karakteristikat e nivelit të lartë dhe avantazhit kryesor të modelit të bazuar në graf, të konkluzioneve dhe të mësuarit global. Parseri përcakton një pemë fillestare parsimi duke modifikuar parashikimet bazuar në një faktorizim të nivelit të parë. Për të realizuar këtë gjë përdoret një rrjet CNN. Më pas karakteristikat e niveleve të larta, si gjyshërit, vëllezërit apo xhaxhallarët, përcaktohen mbi këtë pemë fillestare dhe përdoren për të përcaktuar në mënyrë iterative pemën përfundimtare të parsimit. Teorikisht procesi iterativ do të ekzekutohet deri sa të mos identifikohen ndryshime të reja, por praktikisht eksperimentet e realizuara përcaktojnë që nuk është e nevojshme të realizohen më shumë se dy rifreskime. Vlerësimet eksperimentale tregojnë se ky parser ka saktësi më të mirë se ato ekzistuesit (Zheng, 2017).

Vitet e fundit janë zhvilluar shumë sisteme *pipeline* që realizojnë segmentim dhe tokenizim të fjalëve në fjali, etiketimin me pjesë të ligjëratës dhe karakteristika morfologjike, temëzim dhe parsim. Këtu mund të përmendim OpenNLP dhe NLTL.

UDPipe është një tjetër sistem pipeline që realizon të gjitha funksionalitetet e përmendura më sipër. Qëllimi i zhvilluesve të UDPipe ishte të propozonin në sistem i cili është i thjeshtë në përdorim dhe për personat me njohuri jo të avancuara në gjuhësi apo programim, të pavarur nga gjuha, që nuk përdor fjalor morfologjik apo sintaksor, në një tokenizer të paratrajnuar, me licencë të hapur dhe me saktësi të lartë. Ky sistem bazohet në një rrjet neural të të mësuarit e thelluar LSTM (Straka & Strakova, 2017).

Një tjetër sistem pipeline i bazuar në të mësuarin e thelluar që realizon segmentimin e fjalive dhe tokenizim, etiketimin e pjesëve të ligjëratës dhe karakteristikave morfologjike, temëzim dhe parsim është dhe parseri Turk Neural Pipeline Parser që është trajtuar me detaje në këtë disertacion në çështjen 5.7 . Ky sistem është përdorur për të propozuar në këtë disertacion një etiketues morfologjik dhe temëzues për gjuhën shqipe.

Stanford është një tjetër sistem pipeline i bazuar në rrjet neural që realizon segmentim, tokenizim, etiketim të pjesëve të ligjëratës dhe parsim. Ky sistem ka saktësi të lartë kur përdoret në korpuse të mëdhenj. Për të përmirësuar qëndrueshmërinë në këtë sistem është propozuar përdorimi i një metodë të re të kombinimit të njohurive statistikore me rrjete neurale fleksibël dhe të fuqishëm. Për të siguruar një parashikim të qëndrueshëm është propozuar përdorimi i një klasifikuesi *biaffine* për parashikimin e përbashkët të etiketës së pjesëve të ligjëratës dhe të etiketave të karakteristikave morfologjike. Temëzuesi përdor një klasifikues redaktimi që përmirëson qëndrueshmërinë e tij (Qi, et al., 2018).

Përveç metodave të të mësuarit të kontrolluar që kemi diskutuar më lartë, në shumë punime kërkimore janë propozuar sisteme të ndryshme etiketimi morfologjik, temëzimi apo rrënjëzimi dhe parsimi të bazuara në të mësuarit e pakontrolluar. Në punimin përmbledhës të (Christodoulopoulos, et al., 2010) janë vlerësuar dhe krahasuar disa nga këto metoda ekzistuese. Vlerësimi krahasues ka treguar se disa nga metodat e vjetra tepër të thjeshta kanë ende saktësi të lartë në etiketim në krahasim me metodat e propozuar më vonë.

5.3. Teknikat e gjetjes së rrënjës dhe temës së fjalës

Teknikat e para të përdorura për gjetjen e rrënjës apo temës së fjalës janë të bazuara në rregulla dhe të implementuara për një gjuhë të caktuar. Për t'i përdorur këto sisteme në gjuhë të tjera do të duhet të riprogramoheshin sipas rregullave të gjuhës.

Më 1968, Lovin zhvilloi algoritmin e parë të bazuar në rregulla për gjetjen e rrënjës së fjalës për gjuhën angleze. Ky algoritëm bazohet në eliminimin e prapashtesave më të gjata për të nxjerrë rrënjën e një fjale. Algoritmi ka dy faza: eliminimi i skajeve dhe trajtimi i rrënjës së mbetur. Në fazën e parë fshihen nga fjala prapashtesa më e gjatë e fjalës që përputhet më një listë të parapërcaktuar prej 294 prapashtesash. Secila nga këto prapashtesa është lidhur me një nga 29 instruksionet e kushtëzuara të përcaktuara në algoritëm. Më tej

në fazën e dytë, pjesa e fjalës e nxjerrë nga faza e parë trajtohet për të rregulluar disa përjashtime gramatikore, si p.sh bashkëtingëlloret e dyfishta ose shumësi i formuar në mënyrë të parregullt, duke përdorur një nga 35 rregullat e transformimit të përcaktuar në algoritëm. Ky rrënjëzues është shumë i shpejtë por ka një numër të madh rregullash. Gjithashtu algoritmi mund të ketë dhe një numër të madh gabimesh duke qenë se në të është përcaktuar një listë e limituar prapashtesash të cilat mund të mos përfshijnë të gjithë listën e prapashtesave që përdoren në gjuhën angleze (Lovins, 1968).

Rrënjëzuesi i Dawson është një përmirësim i rrënjëzuesit së Lovin bazuar në dy modifikime: në një listë të zgjeruar prapashtesash prej 1200 prapashtesash dhe përdorimi i një teknike të pjesshme të përputhjes, që përputh rrënjët që kanë një nivel të caktuar përputhjeje. Përdorimi i një liste të madhe prapashtesash bën që algoritmi të kërkoj kohë të gjatë për nxjerrjen e rrënjës dhe kërkon shumë memorie për t'u ekzekutuar (Dawson, 1974).

Një nga algoritmet për gjetjen e rrënjës së fjalës më i përdorshëm në GJIK është rrënjëzuesi i Porter-it. Ky është një algoritëm i bazuar në rregulla, i implementuar në pesë hapa, ku janë përcaktuar rreth 60 prapashtesa, dy rregulla për regjistrimin dhe një për varësinë e kontekstit për të përcaktuar nëse një prapashtesë duhet fshirë apo jo. Fillimisht algoritmi është implementuar për gjetjen e rrënjëve të fjalëve në gjuhën angleze dhe më vonë u implementua dhe për gjuhë të tjera si gjuha gjermane, frënge, ruse, etj. (Porter, 1980). Më 2001, Porter krijoj platformën Snowball (Porter, 2001), një version i përmirësuar i algoritmit të parë dhe përfshin implementime për gjuhë të ndryshme. Rrënjëzuesi ka saktësi të mirë në gjetjen e rrënjës por është tepër kompleks në terma të instruksioneve që duhet të ekzekutohen.

Mayfield dhe McNamee (2003) propozojnë përdorimin e n-gram për përcaktimin e rrënjës së fjalës për sisteme të pavarura nga gjuha. Vlerësimet eksperimentale për gjuhë të ndryshme tregojnë që përdorimi i një n-gram të vetëm si rrënjë e një fjale është një metodë shumë efektive.

Sistemet e propozuara vitet e fundit bazohen në përdorimin e të mësuarit e automatizuar, të rrjeteve neurale dhe të mësuarit e thelluar. Këto sisteme kanë saktësi më të lartë në gjetjen e rrënjës apo temës së fjalës në krahasim me sistemet e bazuar në rregulla. Në ndryshim nga sistemet e bazuara në rregulla këto sisteme kanë nevojë për një korpus trajnimi nga i cili modeli mund të mësojë për të parashikuar rrënjën apo temën e fjalës të teksteve të panjohur. Sisteme të tilla bazohen në një model editimi të pemës së klasifikimit, ku pemët e editimit apo rregullat e transformimit nga fjala tek tema e fjalës nxirren nga të dhënat që përdoren për trajnim dhe më pas një klasifikues trajnohet për gjetjen e temës së saktë të fjalës, ose në një model sekuencë-pas-sekuence ku modeli trajnohet për të gjetur temën e fjalës karakter pas karakteri.

Sistemet e temëzimit apo rrënjëzimit mund të jenë sisteme të cilat marrin parasysh kontekstin në të cilin është përdorur fjala në fjali ose që nuk e marrin atë parasysh. Marrja parasysh e kontekstit të përdorimit të fjalës në fjali ndihmon në parashikime të sakta të

temës ose rrënjës në rastin kur kemi fjalë të dykuptimshme. Në këto sisteme, etiketat e pjesëve të ligjëratës dhe të karakteristikave morfologjike japin informacionin e duhur mbi kontekstin e përdorimit të fjalës.

Temëzuesi i sistemit pipeline UDPipe është një temëzues i bazuar në modelin pemë. Në këtë tip temëzues fjala nuk duhet të njihet paraprakisht duke qenë se rregulli i gjenerimit të temës njihet gjatë fazës së trajnimit. Procesi i temëzimit në këtë sistem realizohet veçmas procesit të etiketimit të pjesëve të ligjëratës për të patur një saktësi më të mirë. Ky temëzues është përdorur për rreth 50 gjuhë të ndryshme dhe ka saktësi të mirë (Straka & Strakova, 2017).

Temëzuesi Lematus është një temëzues i cili në gjetjen e temës së fjalës merr parasysh kontekstin e përdorimit të saj duke përdorur etiketat e pjesëve të ligjëratës dhe karakteristikave morfologjike. Qëllimi pse është propozuar një temëzues i tillë është në përcaktimin sa më saktë të temës së fjalëve të pavëzhguara më parë nga temëzuesi dhe të fjalëve të dykuptimshme. Temëzuesi bazohet në një arkitekturë enkoder-dekoder i bazuar në nivel karakteri. Lematus është vlerësuar eksperimentalisht në 20 gjuhë duke përdorur si implementimin e versionit që merr parasysh kontekstin e fjalës dhe versioni që nuk e merr parasysh atë. Rezultatet tregojnë se konteksti ndihmon më shumë në fjalët e dykuptimshme se sa në fjalët e pavëzhguara më parë nga modeli (Bergmanis & Goldwater, 2018).

Temëzuesi i sistemit Turku Neural Parser Pipeline (Kanerva, et al., 2020) bazohet në një model sekuencë-pas-sekuence. Ky temëzues është trajtuar në çështjen 5.7.3 dhe është përdorur në këtë disertacion për propozimin e një temëzuesi për gjuhën shqipe.

5.4. Përmbledhje e literaturës për gjuhën shqipe

Në këtë pjesë, kemi realizuar një përmbledhje të punimeve shkencore të cilat kanë propozuar mjete për etiketimin morfologjik dhe sintaksor në gjuhën shqipe dhe për gjetjen e temës apo fjalëve në gjuhën shqipe. Gjuha shqipe karakterizohet nga një gramatikë komplekse dhe nga një sistem i pasur lakimi të fjalëve që e bën sfidues procesin e etiketimit morfologjik dhe sintaksor dhe gjetjen e temës apo të rrënjës së fjalës. Në vitet e fundit janë bërë disa përpjekje për të zhvilluar mjete të cilat të përdoren për etiketimin morfologjik në gjuhën shqipe të cilat do t'i trajtojmë më në detaje në vazhdim. Për t'u theksuar është se asnjë nga këto mjete aktualisht nuk është i përdorshëm, për arsye sepse një pjesë e mjeteve bazohen në korpuse shumë të vogla të etiketuara dhe nuk kanë performancë të mirë ose sepse nuk janë të publikuar nga zhvilluesit për t'u përdorur gjerësisht. Në Tabelën 5.1 paraqitet një përmbledhje e këtyre punimeve.

Mund të themi që Trommer dhe Kallulli (2004) ishin të parët që në punimin e tyre propozuan një etiketues morfologjik të bazuar në rregullat për shqipen standarde. Etiketuesi ka dy komponentë: një tokënizues dhe një analizues të thjeshtë morfologjik. Tokëniuesi është një skript i thjeshtë në Python. Analizuesi morfologjik ka tre komponentë: një leksik operativ, një grup rregullash morfologjike dhe një interpretues të rregullave. Në

Tabela 5.1 Përmbledhje e punimeve për zhvillimin e mjeteve të përpunimit morfologjik për gjuhën shqipe

PUNIMI	VITI	MJETI	KORPUSI	NR. TOKEN	SAKTËSIA	NR. ETIKETA/ RREGULLA	SHKARKUESHËM
(Trommer & Kallulli, 2004)	2004	Etiketues morfologjik	Novela Kadaresë dhe artikuj gazeta	1000	precizioni 96% - 98%. recall: 92%-95%	17 etiketa	Po
(Piton & Lagji, 2007) dhe (Piton, Lagji, & Përnaska, 2007)	2007	Fjalor elektronik dhe Finite State Transducers	Fjalor Shqip-Frëngjisht	4951	__	__	Jo
(Salavaçi & Biba, 2012)	2012	Etiketues i PL	Novela, gazeta, libra akademikë	10000	60%	100, 150,200 etiketa	Jo
(Arkhangelskiy, Belyaev, & Vydrin, 2012)		Etiketues PL dhe temëzues	Listë fjalës të etiektuar	125.500	__	62 etiketa	Po
(Morozova & Rusakov, 2014)		Korpus i etiketuar	romane, revista, zyrtarë fetarë dhe tekste shkencorë	31.12 milion		62 etiketa	online për kërkim
(Kadriu, 2013)	2013	Etiketues POS dhe temëzues	Jo korpus trajnimi. Korpusi i testimit 30 artikuj.	32000 fjalë	80%- 93%.	22 etiketa	Jo
(Kirov, et al., 2018)	2018	Analizues morfologjik dhe lematizues	Wikipedia	589tema, 33483 fjalë	__	Skema UniMorph	Po
(Kabashi & Proisl, 2018)	2018	Etiketues POS	fjali	31584	85.96% - 95.10%	79 etiketa	Jo
(Karanikolas, 2009)	2014	rrënjëzues me rregulla	15dokumenta tekst	500	80%		Jo
(Sadiku & Biba, 2012)	2012	rrënjëzues me rregulla	__	__	__	134 regulla	Jo
(Biba & Gjati, 2014)	2014	rrënjëzues me rregulla	__	__	__	__	Jo

analizuesin morfologjik janë përcaktuar 340 rregulla morfologjike që tregojnë lidhjen midis tekstit në hyrje dhe formës së derivuar në dalje. Një rregull paraqitet në formën <left_context, remove, add; lexicon_category; tag >, ku where left_context dhe remove janë shprehje të rregullta dhe add, lexicon_category dhe tag janë stringa. Interpretuesi i rregullave përdor këto çifte rregullash për të nxjerrë informacion mbi formën e fjalës. Skema e etiketimit përbëhet nga një set çiftesh atribut-vlerë të përcaktuara sipas standardit EAGLE të adaptuara për gramatikën e gjuhës shqipe. Skema e etiketimit përmbajnë 17 etiketa: n, v, a, prsp, pjesë, reflp, posp, demp, indp, intp, relp, pa, prep, adv, ptl, seq, conj. Etiketuesi është vlerësuar duke përdorur dy korpuse të vegjël, secili me nga 500 tokens (fjalë) dhe nuk mund të përcaktojmë nëse ky mjet është një mjet i mirë.

Piton et al, (2007) dhe Piton dhe Lagji (2007), në punimet e tyre, kanë propozuar një fjalor elektronik për përpunimin automatik të tekstit në gjuhën shqipe duke përdorur një Finite State Transducrs me NooJ's graphs. Analizimi realizohet në nivel fjale dhe jo në nivel fjalie apo teksti në tërësi duke i dhënë një fokus të veçantë fjalëve të përbëra që nuk gjenden në fjalor. Në ndërtimin e formave të zgjedhuara të fjalëve në gjuhën shqipe është përdorur një fjalor shqip-frëngjisht me 4951 fjalë. Në punim nuk janë specifikuar etiketat e përdorura, por nga pjesa teorike e gjuhës shqipe e theksuar në punim, mund të themi se etiketat e përdorura janë emër, folje (vetore dhe jo vetore), mbiemër, parafjalë, ndajfolje, pasthirrmë, numrat kardinalë dhe rendorë. Një rëndësi të veçantë i është kushtuar fjalëve të përbëra të tipit: fjalë XY të krijuara nga një emër me një numër, fjalët XY të krijuara me prapashtesa, fjalë XY të krijuara nga bashkimi i thjeshtë i dy fjalëve dhe fjalët X-Y të krijuara nga bashkimi i dy fjalëve me një vizë në mes. Autorët nuk e kanë vlerësuar eksperimentalisht sistemin e propozuar.

Etiketuesi i pjesëve të ligjëratës dhe karakteristikave morfologjike i propozuar nga Salavaçi dhe Biba (2012) mund të konsiderohet si etiketuesi i parë statistikor për gjuhën shqipe. Për të ndërtuar këtë etiketues është përdorur OpenNLP. Autorët kanë etiketuar manualisht një korpus prej 10.000 fjalësh duke përdorur disa skema etiketimi, një skemë të ngushtë me 100 etiketa , një skemë të mesme me 150 etiketa dhe një skemë të gjerë me 220 etiketa. Skemat e përcaktuara etiketojnë fjalët jo vetëm për kategorinë e pjesës së ligjëratës që i përkasin por edhe karakteristikat morfologjike sipas formës së fjalës së përdorur. Korpus i etiketuar është përdorur për të trajnuar dy modele, një model Maxent dhe një model Perceptron. Saktësia mesatare e të dy modeleve është 60%. Etiketuesi i propozuar është një etiketues statistikor dhe performanca e tij ndikohet përveç të tjerash dhe nga madhësia e korpusit dhe cilësia e etiketitmi të korpusit të përdorur në trajnim.

Korpusi më i madh i mbledhur dhe etiketuar në gjuhën shqipe është "Albanian National Corpus" i realizuar nga grupi i linguistëve të Shën Petërsburgut me rreth 31.12 milion fjalë. Korpusi përmban tekst nga tregime të shkurtra, romane, revista, zyrtarë fetarë dhe tekste shkencorë, etj. Skema e etiketimit e përdorur ka 62 etiketa të cilat përfshijnë etiketat standarde për: foljen, ndajfoljen, parafjalën, mbiemrin, numërorin, formën e shhkurtër të përemrave, lidhëzën, parafjalën, nyjën, pasthirrmën, përemrin dhe etiketat e tjera për rasën,

gjininë, numrin, shquarsinë, mënyrën e foljes, vetën, kohën e foljes, përfaqësimim verbal, diatezën, etj. Në korpus çdo fjalë është etiketuar me rrënjën përkatëse, përkthimin në anglisht, etiketën e pjesëve të ligjëratës dhe karakteristikat morfologjike të leksemës së formës së fjalës, duke mos e lidhur domosdoshmërish me karakteristikat për përdorimin aktual në fjali (Morozova & Rusakov, 2014). Autorët kanë raportuar se duke u bazuar në katër fjalorë të gjuhës shqipe kanë krijuar manualisht skedarët që përmbajnë një listë fjalësh gramatikore me të gjitha informacionet e leksemës dhe paradigmën e lakimit. Korpusi është etiketuar automatikisht duke përdorur analizuesin morfologjik UniParser, i cili çdo fjale në korpus i cakton në mënyrë të veçuar një etiketë duke marrë parasysh informacionin listën e fjalëve gramatikore të krijuar (Arkhangelskiy, et al., 2012). Procesi i etiketimit nuk merr parasysh kontekstin aktual të formës sintaksore të fjalës në përcaktimin e karakteristikave gramatikore. Autorët raportojnë që recall i modelit të trajnuar për etiketim në gjuhën shqipe është 93%. Parseri dhe fjalorët e etiketuarë janë publikuar online dhe mund të shkarkohen pa pagesë, kurse korpusi i etiketuar nuk mund të shkarkohet por vetëm të përdoret për qëllime kërkimi online.

Kadriu (2013) ka propozuar një etiketues morfologjik duke përdorur paketën NLTK. Ky etiketues për të përcaktuar etiketën e pjesëve të ligjëratës në një tekst të ri përdor një korpus të etiketuar prej 32000 fjalësh dhe një sërë rregullash të implementuara nëpërmjet shprehjeve të rregullta. Etiketuesi është i përbërë nga 6 faza. Në fazën e parë është implementuar një tokenizues i cili segmenton dhe tokenizon tekstin që do të etiketohet. Më tej një lematizues përcakton vetëm temat e emrave dhe foljeve. Lematizuesi bazohet në rregulla për fshirjen e parashtesave dhe prapashtesave të fjalës për të përcaktuar temën e fjalës. Më pas fjala etiketohet me etiketën e pjesëve të ligjëratës duke përdorur NLTK për etiketimin në kaskadë me bllokim. Fillimisht, fjala etiketohet duke përdorur një etiketues bazë Unigram të bazuar në fjalor. Më tej përdoret një etiketues i bazuar në shprehje të rregullta i cili përcakton etiketën e fjalës duke përdorur një listë shprehjesh të rregullta dhe një listë prapashtesash. Më tej teksti etiketohet nëpërmjet një etiektuesi me bllokim UnigrammTagger. Nëse gjendet një fjalë qe nuk është në fjalor apo që nuk mund të gjenerohet nga shprehjet e rregullta, etiketohet si “None”. Më tej moduli i lemma_inverse i kthen këto fjalë në formën e tyre fillestare dhe gjenerohet një fjalor i ri. Ky fjalor i ripërdoret në etiektuesin e dytë UnigrammTagger2 i cili përdor RegexpTageger si. Ky etiketues ka RegexpTagger si backoff. Në këtë mënyrë teksti ri etiketohet me UnigrammTagger2 që ka si etiektues bazë RegexpTagger. Skema e etiektimit ka 22 etiketa. Modeli i etiketuesit të propozuar ka një saktësi deri në 90%, por korpusi i përdorur për të trajnuar dhe testuar modelin është i vogël, dhe etiketuesi gjeneron një numër të konsiderueshëm fjalësh të paetiketuara si fjalët që nuk gjenden në fjalor (Kadriu, 2013).

Një korpus i vogël fjalësh i etiketuar me pjesët e ligjëratës, karakteristikat morfologjike dhe temën është krijuar nga Kirov et al. (2018) nën projektin UniMorph. Korpusi përmban 589 lema dhe 33483 forma fjalësh në total të etiketuara sipas skemës së etiketimit UniMorph. Ky korpus është përdorur në detyrën CoNLL-SIGMORPHON 2017 shared

task për të trajnuar një model etiketuesi dhe si modeli i trajnuar dhe korpusi mund të shkarkohet dhe të përdoret pa pagesë.

Kabashi dhe Proisl (2018) kanë propozuar një korpus të etiketuar morfologjisht për gjuhën shqipe me rreth 2020 fjali, 31584 token-sa të etiketuara manualisht nga dy shqipfolës nativë. Ky korpus është version i përmirësuar i korpusit të propozuar në punimin (Kabashi & Proisl, 2016). Autorët kanë propozuar tri skema etiketimi për gjuhën shqipe, skema a parë e etiketimit është specifikuar nga vetë autorët bazuar në gramatikën e gjuhës shqipe, skema e dytë e etiketimit është adaptim i skemës së parë sipas standardit Google UPOS dhe skema e tretë e etiketimit është adaptim i skenës së parë sipas standardit Universal Dependencies. Skema e parë e etiketimi ka gjithsej 79 etiketa të klasifikuara në 16 kategori kryesore të etiketave: 4 etiketa për emrin, 14 etiketa për foljen, 5 etiketa për mbiemrin, 3 etiketat për ndajfoljen, 14 etiketat për përemrin, 1 etiketë për parafjalën, 6 etiketa për lidhëzën, 2 etiketa për numërorin, 19 etiketa për pjesëzën, 1 etiketë për pasthirrmën, 1 etiketë për shenjat emotive, 3 etiketa për përemrin pronor, 2 etiketa për shkurtimet, 2 etiketa për shenjat e pikësimit, dhe 1 etiketë për elementët jo gjuhësor. Skema e propozuar e etiketimit nuk përshin një numër të madh etiketash si për numrin, gjininë, rasën, etj. Duke u bazuar në gramatikën e gjuhës shqipe, autorët kanë përfshirë në këtë skemë një numër të madh etiketash për të dalluar emrat, mbiemrat dhe ndajfoljet e nyjshëm ose jo. Korpusi i etiketuar është përdorur për të trajnuar dhe testuar pesë modele etiketuesish të bazuar në etiketuesin SoMeWeTa, HMM-based HunPos, TreeTagger, Stanger POS Tagger OpenNLP. Saktësia e modeleve varion nga 85,96% në 95,10%, dhe modeli i etiketiusit SoMeWeTa është modeli me performancë më të mirë me saktësi 95.10% duke përdorur skemën e etiketimi Google UPOS dhe 91.00% duke përdorur skemën e etiketimit të propozuar nga autorët. Si korpusi i etiketuar si modelet e trajnuara nuk gjenden online dhe nuk mund të shkarkohen për t'u përdorur në detyra të ndryshme të përpunimit të gjuhës natyrale.

Toska et al. (2020) kanë propozuar një korpus prej 60 fjalish të etiketuara me rrënjën e fjalë me etiketat e pjesëve të ligjëratës, karakteristikave morfologjike të formës së fjalës dhe me varësinë sintaksore sipas skemës së Universal Dependencies. Ky korpus është i publikuar online dhe mund të shkarkohet pa pagesë, por duke qenë se përmban një numër të vogël fjalish të etiketuara nuk mund të përdoret në trajnimin e një modeli parsuesi por mund të shërbejë si një skemë bazë për gjuhën shqipe.

Tentativa e parë për të zhvilluar një algoritëm të bazuar në rregulla për gjetjen e rrënjës së fjalëve në gjuhën shqipe është bërë nga një jo-folës i gjuhës shqipe, Karanikolas (2009). Ky algoritëm bazohet në parimin e fshirjes së prapashtesës më të gjatë të mundshme nga një fjalë për të gjeneruar rrënjën e saj. Duke u bazuar në 5 libra gramatikorë të gjuhës shqipe autori ka krijuar një listë me 470 lidhëza dhe parafjalë të cilat nuk merren parasysh në algoritëm për t'i gjetur rrënjën duke qenë se janë fjalë të pandryshueshme. Algoritmi i propozuar gjeneron rrënjën e fjalës në tre hapa. Në hapin e parë hiqet prapashtesa më e gjatë që përputhet me fundin e fjalës, më pas fjala e gjeneruar përdoret në hapin e dytë për

të hequr një tjetër prapashtesë më të gjatë të mundshme që përputhet me fundin e fjalës së gjeneruar nga faza e parë. Në këtë fazë zbatohet kushti që pjesa e gjeneruar të jetë në përputhje me modelin VCVCVC, ku V është një ose një sekuencë zanoresh dhe C një ose një sekuencë bashkëtingëlloresh. Në hapin e tretë dhe të fundit në fjalën e gjeneruar nga hapi i dytë ruhet vetëm bashkëtingëllorja e parë e një grupi bashkëtingëlloresh të cilat mund të gjenden në fjalën e gjeneruar nga faza e dytë (Karanikolas, 2014). Ky algoritëm është vlerësuar manualisht në një numër të vogël fjalësh në gjuhën shqipe. Për këtë qëllim është mbledhur tekst nga 15 dokumente në gjuhën shqipe, nga të cilat janë fshirë lidhëzat dhe parafjalët dhe përsëritjet e të njëjtës fjalë. Duke gjeneruar në këtë mënyrë një korpus me 5000 fjalë nga të cilat janë zgjedhur 500 fjalë në mënyrë rastësore për t'u përdorur për gjetjen e rrënjëve nëpërmjet algoritmit. Për të vlerësuar saktësinë e algoritmit, vlerësuesit kanë përdorur një fjalor shqip-greqisht për të përcaktuar nëse rrënja është përcaktuar saktë nga algoritmi. Kjo për arsye se si autori i algoritmit si vlerësuesit nuk janë njohës të gjuhës shqipe. Saktësia e raportuar për këtë algoritëm është 80%, gjithsesi numri i vogël i fjalëve të përdorura për testim dhe fakti që autorët nuk janë njohës të gjuhës shqipe nuk mund të themi nëse është një algoritëm i mirë për t'u përdorur. Gjithashtu ekspertët theksojnë se në algoritëm nuk janë marrë parasysh të gjitha rregullat e krijimit të fjalëve duke u nisur nga rrënja e tyre. Në Karanikolas (2013) përshkruhet me detaje formati i të dhënave të përdorura. Nga kërkimet e realizuara nuk rezulton që algoritmi dhe korpusi të jetë i publikuar online.

Algoritmi JStem i propozuar nga Sadiku dhe Biba (2012) mund të konsiderohet si algoritmi i parë i bazuar në rregulla i zhvilluar nga folës të gjuhës shqipe për të gjetur rrënjët e fjalëve. Ky algoritëm është i implementuar në Java dhe bazohet në rregullat e formimit të fjalëve në gjuhën shqipe. Algoritmi ka 5 hapa, secili i përbërë nga një set rregullash. Në total në algoritëm janë përcaktuar 134 rregulla për të fshirë prapashtesat dhe parashtesat e fjalëve në gjenerimin e rrënjës së saj. Në algoritëm është përcaktuar dhe një listë parafjalësh dhe lidhëzash të cilat nuk merren parasysh në gjenerimin e rrënjës duke qenë se janë fjalë të pandryshueshme. Rregullat janë të implementuara si kushte if-else dhe në një hap të algoritmit mund të ekzekutohet vetëm një rregull. Algoritmi fshin nga fjala parashtesën dhe prapashtesën më të gjatë të mundshme me kusht që rrënja e gjeneruar të mos jetë një fjalë me më pak se dy karaktere. Në disa raste rregullat e fshirjes së parashtesave dhe prapashtesave aplikohen disa herë. Në këtë algoritëm nuk janë specifikuar rregulla për formimin e shumësit, formën gjinore femërore, mashkullore dhe neutrale. Algoritmi i propozuar nuk është vlerësuar për saktësinë e gjetjes së rrënjës së fjalëve por është përdorur për parapërpunim teksti në detyrën e klasifikimit të tekstit sipas temave. Autorët theksojnë se aplikimi i algoritmit në këtë detyrë rrit performancën e modelit të trajnuar për klasifikim.

Algoritmi JStem është zgjeruar nga Biba dhe Gjati (2014) për të gjeneruar rrënjën e fjalëve të përbëra në gjuhën shqipe. Duke analizuar strukturën morfologjike të fjalëve të përbëra në gjuhën shqipe, autorët kanë propozuar një set rregullash për gjenerimin e

rrënjëve të fjalëve të përbëra. Dhe në këtë rast çdo rregull është implementuar si një kusht if-else. Rregullat e implementuara në algoritëm marrin në konsideratë fjalët e përbëra të formuara nga dy fjalë të përngjitura, nga dy fjalë të ndara me një vizë ndarëse në mes, me parashtesa, me numërorë, nga një emër i lidhur me një folje ose mbiemër. Fillimisht algoritmi i ndan fjalët e përbëra në pjesët e tyre përbërëse dhe më pas gjen rrënjën e secilës pjesë.

5.5. Skema Universal Dependencies

Universal Dependencies (UD) është një platformë e zhvilluar për një komunitet të gjerë për të krijuar banka të dhënash linguistike të etiketuara për gjuhë të ndryshme duke u bazuar në një kornizë leksikologe të bazuar në vartësi. Skema e etiketimi është zhvilluar mbi bazën e tri skemave: Stanford dependencies, Google universal part-of-speech tags dhe Interset interlingual for morphosyntatic tagsets. Në këtë platformë etiketimi konsiston në segmentimin/tokenizimin e fjalëve, përcaktimin e temës së fjalës, përcaktimin e etiketës së pjesëve të ligjëratës, përcaktimin e etiketave të karakteristikave morfologjike dhe në analizën sintaksore duke u bazuar në marrëdhënien sintaksore midis entiteteve/fjalëve në fjali (Nivre, et al., 2020).

Qëllimi i UD është të sigurohet një skemë sa më qëndrueshme për etiketimin linguistik të tekstit dhe të krijohen banka të dhënash linguistike të etiketuara në të gjitha gjuhët të cilat mund të shërbejnë në aplikime të ndryshme. Skema e etiketimit UD bazohet në një listë etiketash të pjesës së ligjëratës të mirë përcaktuar dhe që nuk janë të ndryshueshme dhe një listë të hapur për karakteristikat morfologjike të cilat mund të zgjerohen për të përfshirë karakteristika specifike të gjuhës dhe një listë etiketash që përfaqësojnë lidhjet sintaksore midis fjalëve në fjali e cila mund të zgjerohet për të përfshirë lidhje specifike sipas gjuhës. Në skemën UD etiketimi bazohet në nivel fjale sintaksore dhe një rresht etiketim në këtë skemë i përket vetëm një fjale sintaksore. Të dhënat e etiketuara sipas skemës UD ruhen në një file të tipit të veçantë CoNLL-X dhe pikërisht versioni CoNLL-U i cili shpjegohet me detaje në çështjen 5.6.

Në Tabelën 5.2 paraqitet lista e të gjithë etiketave që përshin skema UD për etiketimin e pjesëve të ligjëratës, karakteristikat morfologjike dhe lidhjet sintaksore.

5.5.1. Segmentimi/tokenizimi i fjalëve

Në skemën e etiketimit UD, segmentimi i fjalive dhe tokenizimi është një proces shumë i rëndësishëm duke qenë se etiketimi dhe përcaktimi i varësive sintaksore realizohet në nivel fjale sintaksore. Pra fjala nuk ndahet në morfema trajtëformuese dhe më pas të etiketohet. Në rastin kur kemi një fjalë klitike kjo fjalë duhet të ndahet në fjalët sintaksore nga e cila është formuar dhe të etiketohet secila fjalë më vete. P.sh. trajta e shkurtër e përemrave vetor m'i = më+i do të duhet të ndahet dhe do të kemi etiketim të veçantë të

“më” dhe të “i”. Në këtë rast një token konsiderohet si një token shumëfjalësh ku një fjalë e vetme i korrespondon disa fjalëve sintaksore.

Në versionin UD v2 të skemës UD lejohen etiketimi si një fjalë e vetme dhe për fjalët që mund të përmbajnë një hapësirë midis në rastin e disa gjuhëve të veçanta kur një fjalë mund të shkruhet e ndarë në rrokje. Gjithashtu hapësira është e lejueshme dhe në shkrimin e numrave në formatin 10 000 apo shkurtimeve p. sh.. Të dyja këto raste nëse përdoren duhet të përcaktohen në manualin e gjuhës përkatëse. Bazuar në rregullat e gjuhës shqipe ne mund të kemi përdorimin e hapësirës vetëm në paraqitjen e numrave.

Në skemën UD nëse në një fjali dy token-a (që zakonisht mund të jenë një fjalë dhe një shenjë pikësimi) nuk duhet të jenë të ndara me një hapësire midis tyre, në rreshtin e etiketimit të fjalës/token-it të parë në fushën përkatëse të file-it CoNLL-U do të përdoret atributi *AfterSpace=No*. Kjo për të përcaktuar që në shkrimin e fjalisë nuk ka një hapësirë pas këtij token dhe tokenit pasardhës (Nivre, et al., 2020).

Tabela 5.2 Lista e etiketave të skemës UD

Etiketa e pjesëve të ligjëratës	**Etiketat e karakteristikave morfologjike**		**Marrëdhënia sintaksore**		
			Klauzolare		
	Lakueshëm	**Leksikore**	**Kryesore**	**Jo kryesore**	**Nominale**
ADJ	Animacy	Abbr	nsubj	advcl	acl
ADP	Aspect	Foreign	csubj	advmod	amood
ADV	Case	NumType	ccomp	aux	appos
AUX	Clusivity	Poss	iobj	cop	case
CCONJ	Definite	PronType	obj	discourse	clf
DET	Degree	Reflex	xcomp	dislocated	det
INTJ	Evident	Typo		expl	nmod
NOUN	Gender			mark	nummod
NUM	Mood			obl	
PART	NounClass			vocative	
PRON	Number		**Lidhja**	**MWE**	**Speciale**
PROPN	Personal		cc	compound	dep
PUNCT	Polarity		conj	fixed	goeswith
SCONJ	Polite		list	flat	orphan
SYM	Tense		parataxis		punct
VERB	VerbForm				reparandum
X	Voice				root

5.5.2. Etiketimi morfologjik

Në skemën UD etiketimi morfologjik i një fjale sintaksore përbëhet nga tri nivele:

1. Tema e fjalës që është forma bazë e fjalës që paraqitet në fjalor;
2. Etiketa e pjesëve të ligjëratës që përfaqëson kategorinë gramatikore që i përket fjala;
3. Etiketa/t e karakteristikave morfologjike që përcaktojnë karakteristikat morfologjike të kësaj forme specifike të fjalës.

Skema UD lejon që nëse:

- në temën e një fjale nuk është përcaktuar në etiketim në fushën përkatëse të shënohet shenja "_";
- në fushën e etiketës të pjesëve të ligjëratës nuk duam të përcaktojmë një vlerë do të duhet të shënojmë "X";
- në fushën e etiketës së karakteristikave morfologjike mund të mos përcaktohet një vlerë dhe të vendoset shenja "_".

Tema e fjalës duhet të jetë një fjalë e vetme dhe në të nuk lejohen hapësira. Gjithashtu një fjalë mund t'i përkas vetë një pjese së ligjëratës prandaj dhe në fushën e etiketës së pjesëve të ligjëratës duhet të përcaktohet vetëm një etiketë. Në skemën UD janë përcaktuar 17 etiketa të pjesëve të ligjëratës të cilat janë paraqitur në Tabelën 5.2, më tej do të shpjegohet përdorimi i secilës etiketë. Në fushën e etiketave të karakteristikave morfologjike mund të përcaktohet asnjë, një ose disa karakteristika gramatikore të formës përkatëse të fjalës duke përdorur çifte të tipit karakteristike=vlerë të renditura sipas rendit alfabetik të emrit të karakteristikës dhe të ndara me shenjën "|" çiftet midis tyre. Skema lejon që për një karakteristikë të caktuar të përcaktohen dy vlera ku vlerat ndahem ne njëra-tjetrën me ",", karakteristike=vlerë,vlerë. Dhe në këtë etiketë nuk lejohet përdorimi i hapësirave. Për çdo etiketë të pjesës së ligjëratës janë përcaktuar karakteristika të caktuara që mund të përdoren. Për shembull karakteristika PronType që përdoret për të përcaktuar tipin e një përemri nuk mund të përdoret si etiketë në listën e karakteristikave morfologjike të një fjale që është etiketuar me etiketën e pjesës së ligjëratës NOUN, që përdoret për emrin. Karakteristikat morfologjike të përdorura në këtë skemë do të shpjegohen në vazhdim. Skema UD jep lirshmërinë që në varësi të gramatikës së gjuhës të mos përdoren të gjitha etiketat në etiketimin e tekstit në një gjuhë të caktuar.

Etiketat e pjesëve të ligjëratës: përdoren për shënuar kategoritë kryesore të pjesëve të ligjëratës. Më poshtë janë listuar në rend alfabetike të gjitha etiketat dhe për se përdoren:

- ADJ – përdoret për të përcaktuar mbiemrin;
- ADP – përdoret për të përcaktuar parafjalën;
- ADV – përdoret për të përcaktuar ndajfoljen;
- AUX – përdoret për të përcaktuar foljet ndihmese;
- CCONJ – përdoret për të përcaktuar lidhëzat bashkërenditëse;
- DET – përdoret për të përcaktuar nyjet;
- INTJ – përdoret për të përcaktuar pasthirrmat
- NOUN – përdoret për të përcaktuar emrin;

- NUM – përdoret për të përcaktuar numërorin;
- PART – përdoret për të përcaktuar pjesëzën;
- PRON – përdoret për të përcaktuar përemrin;
- PUNCT – përdoret për të përcaktuar shenjat e pikësimit;
- PROPN – përdoret për të përcaktuar emrat e përvetshëm;
- SCONJ – përdoret për të përcaktuar lidhëzat nënrenditëse;
- SYM – përdoret për të përcaktuar simbolet;
- VERB – përdoret për të përcaktuar foljet
- X – përdoret për të përcaktuar çdo fjalë tjetër që nuk mund te përcaktohet me anë të etiketave të tjera ekzistuese.

Lista e karakteristikave në skemën UD është shumë e gjerë, ne do të trajtojmë karakteristikat që janë përdorur dhe në etiketimin e korpusit në gjuhën shqipe. Pra, etiketat dhe vlerat e tyre të cilat mund të përdoren në etiketim duke patur parasysh rregullat gramatikore të gjuhës shqipe.

Abbr

Format e shkurtuara të fjalëve mund t'u përcaktohet tema si tema e plotë e fjalës së pashkurtuar nëse forma e shkurtuar përfaqëson vetëm një fjalë të plotë. Nëse forma e shkurtuar përbëhet nga disa fjalë në formë të gjatë atëherë si temë të kësaj fjale duhet të jetë përsëri forma e shkurtuar. Për shembull nëse kemi shkurtimin p.sh tema e saj duhet të përcaktohet po si p.sh. dhe jo për shembull, kurse nëse kemi shkurtimin angl. tema e saj mund të përcaktohet si anglisht. Në etiketimin e shkurtimeve në fushën e etiketave të karakteristikave morfologjike përveç karakteristikave në varësi të formës së përdorur duhet të shtohet dhe karakteristika morfologjike abbreviation si Abbr=Yes,

AdvType

Përdoret për të përcaktuar llojin e ndajfoljes. Mund të përcaktohet vetëm me fjalë të cilat janë etiketuar me etiketën ADV që përcakton ndajfoljet. Vlerat që mund të marrë duke u bazuar në gramatikën e gjuhës shqipe janë: Man (ndajfolje mënyre), Loc (ndajfolje vendi), Tim (ndajfolje kohe), Deg (ndajfolje sasie), Cau (ndajfolje shkaku/qëllimi).

Case

Përdoret për të përcaktuar rasën e emrave ose të përemrave. Zakonisht përdoret për emrat por në gjuhë të ndryshme përdoret dhe për pjesë të tjera të ligjëratës. Në gjuhën shqipe është përdorur për fjalët e etiketuara si NOUN, PROPN, PRON, ADJ dhe DET. Vlerat që merr sipas gramatikës së gjuhës shqipe janë: Nom (rasa emërore), Gen (rasa gjinore), Dat (rasa dhanore), Acc (rasa kallëzore), Abl (rasa rrjedhore).

Definite

Përdoret për të përcaktuar trajtën e emrave, mbiemrave dhe nyjeve. Në gjuhën shqipe është përdorur për fjalët e etiketuara si NOUN dhe PROPN. Vlerat që merr sipas gramatikës së gjuhës shqipe janë: Definite (trajta e shquar) dhe Indefinite (trajta e pashquar).

Degree

Përdoret për të përcaktuar shkallën e mbiemrit dhe ndajfoljeve. Në gjuhën shqipe është përdorur për fjalët e etiketuara si ADJ dhe ADV për të shprehur vetëm shkallën pohore me vlerë: Pos. Duke u bazuar në gramatikën e gjuhës shqipe qoftë për mbiemrin dhe për ndajfoljen ne kemi tri shkallë. Në skemën UD ekzistojnë dhe dy shkallët e tjera me vlera Cmp (shkalla krahasore) dhe Sup (shkalla sipërore) por duke qenë se në gjuhën shqipe këto dy shkallë formohen nga shprehje dhe jo si një fjalë e vetme, nuk mund të përdoren në etiketim.

Foreign

Përdoret për të përcaktuar një fjalë në një gjuhë të huaj nga gjuha e korpusit që po etiketohet. Merr një vlerë boleane Yes (po) dhe përcakton një fjala është në gjuhë të huaj. Mund të përdoret me çdo etiketë të pjesëve të ligjëratës, madje dhe me vlerën "X" kur nuk është përcaktuar etiketa e pjesëve të ligjëratës.

Gender

Përdoret për te përcaktuar gjininë e emrave dhe pjesëve të tjera të ligjëratës sipas gramatikës së gjuhës në të cilën po etiketohet. Në gjuhën shqipe është përdorur për fjalët e etiketuara si NOUN, PROPN, PRON, ADJ dhe DET. Vlerat që merr sipas gramatikës së gjuhës shqipe janë: Fem (gjinia femërore), Masc (gjinia mashkullore) dhe Neut (gjinia asnjanëse).

NumType

Përdoret për të përcaktuar llojin e numëruesit. Shoqërohet me etiketën NUM dhe mund të marrë vlerat: Card (vlerë numerike) dhe Ord (numër rendor).

PronType

Përdoret për te treguar llojin e përemrit, mbiemrat pronominalë, numërorët pronominalë dhe ndajfoljet pronominale. Në gjuhën shqipe është përdorur për fjalët e etiketuara si PRON për të përcaktuar tipin e përemrit dhe DET për të përcaktuar që është nyje . Vlerat që merr sipas gramatikës së gjuhës shqipe për etiketën PRON janë: Dem (përemër dëftor), Prs (përemër vetor), Int (përmor pyetës), Ind (përemër i pacaktuar), Rel (përmor lidhor) dhe për etiketën DET është Art (nyje).

Poss

Përdoret për te përcaktuar pronësinë dhe ka vlerë boleane. Në gjuhën shqipe është përdorur për të përcaktuat përemrin pronore duke shoqëruar etiketën PRON bashkë me karakteristikën PronType=Prs. Në etiketim përemri pronor përcaktohet si një përemër i tipit vetore dhe që ka cilësinë e pronësisë duke përdorur karakteristikën Poss me vlerë Yes (po).

Reflex

Përdoret për te përcaktuar përemrat vetvetore dhe ka një vlerë boleane. Në gjuhën shqipe është përdorur për të përcaktuar përemrin vetvetor duke shoqëruar etiketën PRON bashkë me karakteristikën PronType=Prs. Në etiketim përemri vetvetor përcaktohet si një përemër i tipit vetore dhe i referohet vetës së parë duke përdorur karakteristikën Reflex me vlerë Yes (po).

Aspect

Përdoret për të përcaktuar aspektin e një kohe të foljes. Mund të përdoret dhe me pjesë të tjera të ligjëratës. Në gjuhën shqipe është përdorur për foljet për të përcaktuar dy kohët e shkuara, kohën e pakryer dhe kohën e kryer të thjeshtë. Vlerat që merr sipas gramatikës së gjuhës shqipe janë: Imp (koha e pakryer), Perf (koha e kryer e thjeshtë).

Mood

Përdoret për të përcaktuar mënyrën e foljes. Vlerat që merr sipas gramatikës së gjuhës shqipe janë: Ind (mënyra dëftore), Imp (mënyra urdhërore), Cnd (mënyra kushtore), Sub (mënyra lidhore), Des (mënyra dëshirore) dhe Adm (mënyra habitore).

Number

Zakonisht përdoret për të përcaktuar numrin e emrave por në varësi të gjuhës mund të përdoret dhe për pjesë të tjera të ligjëratës. Në gjuhën shqipe është përdorur për të përcaktuar numrin e emrave, përemrave, mbiemrave, nyjeve dhe foljeve. Vlerat që merr sipas gramatikës së gjuhës shqipe janë: Sing (numri njëjës) dhe Plur (numri shumës).

Person

Përdoret për të treguar veten e përemrit dhe foljeve. Në gjuhën shqipe është përdorur me etiketat PRON dhe VERB. Vlerat që merr sipas gramatikës së gjuhës shqipe janë: 1, 2, dhe 3.

Tense

Përdoret për të përcaktuar kohën e foljes. Në gjuhën shqipe pavarësisë se ne kemi 10 kohë të ndryshme, kjo karakteristikë është përdorur për të përcaktuar vetëm kohën e tashme dhe kohën e shkuar. Vlerat që merr sipas gramatikës së gjuhës shqipe janë: Past (koha e

shkuar), Pres (koha e tashme). Më në detaje pse janë përdorur vetëm këto dy kohë në etiketim trajtohet në çështjen 6.2.2.

VerbForm

Përdoret për të përcaktuar format e pashtjelluara të foljes. Duke u bazuar në mënyrën si si formohen format e pashtjelluara në gjuhën shqipe është përdorur vetëm vlera: Part (forma pjesore ose mohore). Më në detaje pse është përdorur vetëm kjo vlere në etiketim trajtohet në çështjen 6.2.2.

Voice

Zakonisht përdoret për të përcaktuar diatezën e foljeve por mund të përdoret dhe me pjesë të tjera të ligjëratës si emra, mbiemra dhe ndajfolje në varësi të gjuhës. Në gjuhën shqipe është përdorur vetëm për të përcaktuar diatezën e foljes, pra vetëm me etiketën VERB. Vlerat që merr sipas gramatikës së gjuhës shqipe janë: Act (veprore) dhe Pass (joveprore).

5.6. Formati CoNLL-U

Teksti i etiketuar sipas skemës UD ruhet në një file me formatin CoNLL-X (Buchholz & Marsi, 2006) dhe pikërisht në versionin CoNLL-U. Teksti i etiketuar ruhet në një file të thjeshtë tekst të koduar me UTF-8, normalizuar në NFC. Fjalitë e etiketuara ndahen nga njëra-tjetra duke përdorur një rresht bosh dhe në fund të file-t gjithashtu përdoret një rresht bosh për të përcaktuar përfundimin e fjalisë së fundit të tekstit të etiketuar. Në këtë file mund të kemi tre tipe rreshtash:

- Rresht bosh për ndarjen e fjalive;
- Rresht për komente të cilat duhet të fillojnë me karakterin “#”;
- Rresht etiketimi.

Çdo fjali përbëhet nga një ose disa tokensa. Etiketimi i çdo tokens të fjalisë realizohet në një rresht të veçantë etiketimi që konsiston në 10 fusha të ndara me një karakteret “tab” midis tyre. 10 fushat e etiketimit sipas renditjes në file janë:

1. ID – indeksi i fjalën në fjali. Numërimi i këtij indeksi në një fjali fillon nga numri 1. Në këtë fushë për tokensat që përbëhen vetëm nga një fjalë indeksi është një numër, p.sh, 1, 2, 3, etj. kurse për tokensat që janë të përbëra nga disa fjalë, pra që duhet të ndahen në dy ose më shumë fjalë sintaksore përdoren vargjet e numrave si 1-2 ose 3-5, etj. në varësi se në sa fjalë ndahet. Rasti tipik i përdorimit të vargjeve të indeksimit në gjuhën shqipe janë trajtat e shkurtra të bashkuara të përemrave.
2. FORM – fjala, simboli ose shenja e pikësimit.

3. LEMMA - tema e fjalës. Në raste specifike mund të jetë vendosur dhe rrënja e fjalës kjo në varësi të gjuhës dhe qëllimit të etiketimit. Në korpusin tonë në gjuhën shqipe ne kemi përdorur rrënjën e fjalës. Në këtë fushë nëse nuk përcaktohet një vlerë duhet të vendoset shenjë "_".
4. UPOS - etiketa e pjesëve të ligjëratës. Në këtë fushë nëse nuk përcaktohet një vlerë duhet të vendoset shenja "X".
5. XPOS - etiketa specifike e gjuhës që mund të lidhet edhe me një skemë tjetër etiketimi. Në këtë fushë nëse nuk përcaktohet një vlerë duhet të vendoset shenja "_".
6. FEATS - lista e karakteristikave morfologjike, në formatin emri_karakteristikë=vlera, të renditura në rendin alfabetik sipas emrit të karakteristikës dhe të ndara midis tyre nga shenja "|". Përveç kësaj për një karakteristikë mund të përcaktohet më shumë se një vlerë duke i ndarë vlerat me presje "," në formën e përgjithshme emri_karakteristikë=vlera1,vlera2, vlera3. Duhet të jetë një stringë e vetme, pra nuk lejohen hapësira. Në këtë fushë nëse nuk përcaktohet një vlerë duhet të vendoset shenja "_".
7. HEAD – rrënja e tokenit aktual. Mund të jetë një vlerë ID ose 0 nëse tokeni aktual është rrënja e pemës sintaksore të fjalisë. Në këtë fushë nëse nuk përcaktohet një vlerë duhet të vendoset shenja "_".
8. DEPREL – etiketa e lidhjes sintaksore e skemës UD në varësi me rrënjën. Nëse për tokenin aktual fusha HEAD ka vlerë 0 atëherë në këtë fushë shënohet "root". Në këtë fushë nëse nuk përcaktohet një vlerë duhet të vendoset shenja "_".
9. DEPS - Grafi i varësive në formën e një liste çiftesh: HEAD: DEPREL
10. MISC - çdo lloj shënimi tjetër. Në këtë fushë nëse nuk përcaktohet një vlerë duhet të vendoset shenja "_". Në këtë fushë mund të përdoret vlera "SpaceAfter=No" nëse midis token-it aktual dhe token-it pasardhës nuk kemi një hapësirë. Shembulli tipik është përdorimi i një shenje pikësimi pas një fjale, p.sh. nëse kemi fjalinë "Përshëndetje!" në rreshtin e etiketimit të fjalës përshëndetje duhet në fushën MISC të përcaktohet kjo gjë. Kjo fushë mund të përdoret për ruajtën e çdo informacioni shtesë që nuk përshtatet në ndonjë fushë tjetër, siç janë shënime specifike të gjuhës, çdo informacion në lidhje me nivelet e tjera gjuhësore siç janë diskurset, apo marrëdhëniet e varësisë. Është e nevojshme që të përcaktohet se çfarë është përshirë në këtë fushë në dokumentin e specifikimit të etiketimit të një gjuhe të caktuar. Në këtë fushë ashtu si në fushën FEATS mund të përcaktohen disa karakteristika në formën emër=vlerë të ndarë me shenjën "|" si një stringë e vetme.

Në tabelën 5.3 paraqitet shembulli i etiketimit të një fjalie në këtë format.

Tabela 5.3 Shembull etiketimi në formatin CoNLL-U

ID	FORM	LEMMA	UPOS	XPOS	FEATS	HEAD	DEPREL	DEPS	MISC
1	Vitin	vit	NOUN	_	Case=Acc|Definite=Def|Gender=Masc|Number=Sing	_	_	_	_
2	e	e	DET	_	Case=Acc|Gender=Masc|Number=Sing|PronType=Art	_	_	_	_
3	Ri	ri	ADJ	_	Case=Acc|Gender=Masc|Number=Sing|Degree=Pos	_	_	_	_
4	do	do	PART	_	_	_	_	_	_
5-6	ta	_	_	_	_	_	_	_	_
5	të	të	PART	_	_	_	_	_	_
6	e	ajo	PRON	_	Case=Acc|Gender=Fem|Number=Sing|Person=3|PronType=Prs	_	_	_	_
7	bësh	bëj	VERB	_	Mood=Sub|Number=Sing|Person=2|Tense=Pres|Voice=Act	_	_	_	_
8	këtu	këtu	ADV	_	AdvType=Loc	_	_	_	SpaceAfter=No
9	.	.	PUNCT	_	_	_	_	_	_

Fushat HEAD dhe DEPREL përdoren për të përcaktuar pemën e varësisë së një fjalie, Vlera në fushën DEPREL duhet të jetë një relacion i varësisë universale. Në korpusin e etiketuar për gjuhën shqipe të prezantuar në këtë disertacion nuk është realizuar etiketimi sintaksor dhe fushat HEAD, DEPREL dhe DEPS janë bosh dhe është vendosur karakteri "_".

5.7. Platforma Turku Neural Parser Pipeline

Parseri Turku Neural Parser Pipeline (Turku Pipeline) është propozuar nga Kanerva et al. (2018). Turku Pipelline është një parser pipeline për analizimin e të dhënave tekst të papërpunuara në UD. Parseri Turku Pipeline përbëhet nga komponentë që ekzekutohen njëri pas tjetrit (në *pipeline*): komponenti për segmentim dhe tokenizim, komponenti për etiketimin e pjesëve të ligjëratës e karakteristikave morfologjik dhe parsim dhe komponenti për gjetjen e temës së fjalës. Ky parser nga rezultatet eksperimentale është renditur i pari për saktësinë në gjetjen e rrënjës së fjalës dhe i dyti për saktësinë e etiketimit të pjesëve të ligjëratës dhe pemës së parsimit në detyrën CoNLL-2018 Shared Task. Për të trajnuar një model për një gjuhë të caktuar Turku Pipeline duhet të trajnohet duke përdorur një korpus të etiketuar sipas skemës UD dhe të ruajtur në formatin CoNLL-U. Në vazhdim do të analizohen komponentët e tij.

5.7.1. Komponenti i segmentimit dhe tokenizimit

Qëllimi i këtij komponenti është të segmentojë/ndajë fjalitë në një tekst dhe të tokenizojë tekstin në tokensa, që është procesi i ndarje së fjalëve ose simboleve të tekstit si njësi të veçanta. Në Turku Pipeline komponenti i segmentimit bazohet në komponentin e segmentimit dhe tokenizimit të UDPipe 1.1 Baseline System (Straka & Strakova, 2017).

Dy proceset, segmentimi i fjalive dhe tokenizimi në UDPipe realizohen bashkërisht në të njëjtin moment. Për të realizuar këto procese përdoret një rrjet me një shtresë dydrejtimësh GRU i cili përcakton për çdo karakter në tekst nëse është karakteri i fundit në një fjali, karakteri i fundit në një token ose nuk është karakteri i fundit në një token. Në shumicën e gjuhëve një token nuk mund të ketë një hapësire dhe rrjeti nuk duhet të përcaktojë fundin e një token-i para një hapësire por të mësojë si të përcaktojë token-sa dhe në fjalët e ngjitura si p.sh. *Përshëndetje!* ose trajta e shkurtër *m'i*. Ky komponent është i aftë që të mësojë të ku duhet të ndajë fjalën për të formuar tokensat dhe ku duhet të ndajë fjalitë edhe pa qenë prezentë një hapësirë. Por gjithashtu është i modifikueshëm në rastin e gjuhëve të cilat lejojnë ekzistencën e hapësirës në një token. Sistemi është në gjendje të identifikojë paragrafët dhe fundin e një dokumenti. Për të identifikuar paragrafin përdoret një rresht bosh kurse për të identifikuar fundin e dokumentit duke ruajtur të dhëna të skedarit që po përpunohet.

5.7.2. Komponenti për etiketim morfologjik dhe parsim

Komponenti për etiketimin e pjesëve të ligjëratës, karakteristikave morfologjike dhe gjenerimin pemës së parsimit i përdorur është një version i modifikuar i parserit Stanford's Graph-based Neural Dependency Parser (Dozat, et al., 2017). Parserit Stanford-it përdor vetëm etiketën e pjesëve të ligjëratës dhe etiketën specifike të gjuhës dhe jo etiketat për karakteristikat morfologjike të fjalës. Për këtë arsye modifikimi i realizuar në këtë sistem për të përcaktuar dhe etiketat e karakteristikave morfologjike është bashkimi i etiketës specifike të gjuhës me etiketat e karakteristikave morfologjike si një etiketë e vetme dhe përdorimi i kësaj etikete të gjeneruar. Në dalje të këtij komponenti etiketat e bashkuar do të ndahet në etiketën specifike të gjuhës dhe në etiketat e karakteristikave morfologjike sipas kolonave të formatin CoNLL-U. Pra kjo etiketë e bashkuar përdoret vetëm brenda këtij komponenti për etiketim dhe parsim. Në rastin e modelit të përdorur për gjuhën shqipe ne nuk kemi përcaktuar një etiketë specifike për gjuhën shqipe dhe kjo fushë është bosh.

Komponenti është një klasifikues i shpërndarë në kohë mbi tokens në një fjali që përdor një rrjet dy-drejtimor LSTM. Ai përbëhet nga dy shtresa të veçanta klasifikimi, një për etiketat e pjesëve të ligjëratës dhe një për etiketat e karakteristikave morfologjike. Enkoderi dy-drejtimor ndahet midis të dy këtyre shtresave të klasifikimit. Në mënyrë më të detajuar komponent do të analizohet në nën çështjet në vazhdim.

5.7.2.1. Arkitektura

Arkitektura bazohet në një model të bazuar në graf të implementuar nëpërmjet një rrjeti neural. Arkitektura e parserit paraqitet në Figurën 5.1. Si hyrje e modelit është një sekuencë token-ash dhe etiketat e pjesëve të ligjëratës përkatëse që kalojnë nëpër një rrjeti me disa shtresa dydrejtimore LSTM, BiLSTM. Dalja nga shtresa e fundit të rrjetit LSTM kalon nëpërmjet katër shtresave të veçanta ReLU që krijojnë katër paraqitje specifike në formë vektori: një të fjalës si varësi e rrënjës së pemës së parsimit, një të fjalës si rrënjë e të gjithë nyjeve që varen nga ajo, një të fjalës si varësi e etiketës së saj dhe një të fjalën si rrënjë e etiketës të varësive të saj. Të katër këto paraqitje vektor përdoren në dy klasifikues *biaffine*, ku i pari llogarit një vlerë për çdo token dhe vlera më e lartë e llogaritur përfaqëson rrënjën më të përshtatshme për atë token dhe i dyti llogarit një vlerë për çdo etiketë për një çift token/rrënjë ku vlera më e lartë përfaqëson etiketën më të përshtatshme për harkun nga rrënja tek nyja vartëse.

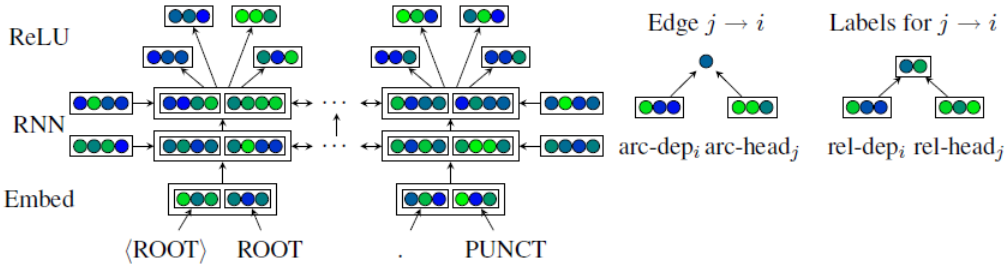


Figura 5.1 Arkitektura e parserit dhe etiketuesit morfologjik (Dozat, et al., 2017)

Arkitektura e etiketuesit të pjesëve të ligjëratës në përgjithësi ka të njëjtën arkitekturë të parserit por me disa ndryshime. Në rrjetin BiLSTM përdoren vektor fjalësh dhe më pas shtresat ReLU përdoren për të krijuar një paraqitje të vetme për çdo tip etikete të pjesëve të ligjëratës. Klasifikuesi *affine* përdoret për çdo tip etikete të pjesëve të ligjëratës.

Në të dy raste, si të parserit dhe të etiketuesit të pjesëve të ligjëratës klasifikuesit trajnohen bashkërisht por parseri dhe etiketuesi i pjesëve të ligjëratës trajnohen më vete (Dozat, et al., 2017).

5.7.2.2. Etiketimi i pjesëve të ligjëratës

Etiketuesi i pjesëve të ligjëratës është një klasifikues *affine* i shpërndarë në kohë përgjatë tokensave të një fjalie. Fillimisht një enkoder që përdor një rrjet njëdrejtimor LSTM i inkorporon tokensat duke përmbledhur së bashku një *learned token embedding*, një *pre-trained token embedding* dhe një *token embedding encoded* nga sekuenca e tij e karaktereve. Më pas këto sekuenca të *embedded tokens* lexohen nga një rrjet dydrejtimor LSTM për të krijuar një paraqitje të përshtatshme sipas kontekstit që është përdorur token-i. Më tej duke përdorur shtresa ReLU të pavarura kjo paraqitje e token-it transformohet për

secilën nga shtresat e klasifikimit *affine*, përkatësisht për UPOS dhe XPOS. Dy shtresat e klasifikimit trajnohen së bashku (Kanerva, et al., 2018).

5.7.2.3. Etiketimi i karakteristikave morfologjike

Për të realizuar etiketimin e karakteristikave morfologjike është përdorur një artificë duke qenë se parseri Stanford (Dozat et al., 2017) nuk e parashikon këtë pjesë. Për këtë qëllim etiketat e karakteristikave morfologjike janë bashkuar me etiketën specifike të gjuhës që përdoret nga parseri Stanford. Në këtë mënyrë etiketa specifike e gjuhës dhe etiketat e karakteristikave morfologjike konsiderohen si një stringë e vetme. Nëse një fjalë ka në fushën e saj XPOS vlerën N dhe në fushën FEATS vlerat Case=Nom|Number=Sing etiketa e bashkuar dhe që përdoret nga sistemi do të jetë XPOS=N| Case=Nom|Number=Sing. Kjo string e përbërë do të parashikohet si një etiketë e vetme nga etiketuesi. Pas realizimit të etiketimit dhe parsimit stringa e përbërë ndahet në dy pjesët nga e cila u formua dhe secila pjesë shkruhet në kolonën e përcaktuar të formatit CoNLL-U. Metoda e përdorur ka rezultuar frytdhënëse dhe ndikon në përmirësimin e saktësisë dhe të parsuesit duke qenë se sistemi merr më shumë informacion (Kanerva, et al., 2018).

5.7.2.4. Pema e parsimit

Për parsim përdoret i njëjti sistem që është përdorur dhe për etiketimin e pjesëve të ligjëratës. Fillimisht një enkoder që përdor një rrjet një-drejtimor LSTM i inkorporon tokensat duke përmbledhur së bashku një *learned token embedding*, një *pre-trained token embedding* dhe një *token embedding encoded* nga sekuenca e tij e karaktereve. Më pas këto sekuenca të *embedded tokens* të cilat janë të shoqëruara nga embeddings përkatëse të etiketës së pjesëve të ligjëratës lexohen nga një rrjet dy-drejtimor LSTM për të krijuar një paraqitje të përshtatshme sipas kontekstit të token-it. Më tej duke përdorur katër shtresa ReLU të pavarura për dy klasifikues të ndryshëm biaffine për të identifikuar lidhjet (HEAD) dhe varësitë (DEPREL). Vetëm parashikimet më të mira dekodohen për të krijuar pemën e parsimit. Këto klasifikues trajnohen së bashku duke përmbledhur humbjen cross-entropy të tyre (Kanerva, et al., 2018).

5.7.3. Komponenti i gjetjes së temës së fjalës

Temëzim është procesi i gjetjes së temës së fjalës, që është forma e fjalës që gjendet në fjalor. Tëmëzuesi është aplikacioni i cili gjen në mënyrë automatike temën e një fjale.

Në gjuhët e pasura me forma morfologjike fjalësh, temëzuesi është një komponent shumë i rëndësishëm për përpunimin e gjuhës natyrale. Kjo bazuar në faktin që në shumë aplikime të metodave të të mësuarit e automatizuar, si nxjerrja e fjalëve kyçe, klasifikim teksti, kërkim në tekst, etj. si hyrje përdoret tema e fjalës dhe jo forma aktuale e fjalës.

Në Turku Pipeline komponenti i gjetjes së temës së fjalës është propozuar nga Kanerva et al. (2020). Ky komponent është një modeli sekuence-pas-sekuence (angl. *sequence-to-*

sequence), që nxjerr temën e fjalës karakter pas karakteri duke përdorur formën e fjalës dhe etiketën e pjesëve të ligjëratës dhe të karakteristikave morfologjike. Duke qenë se ky temëzues nuk përdor vetëm formën e fjalës por dhe etiketat, saktësia e tij është më e lartë duke gjetur temën përkatëse të fjalës në mënyrë më të saktë dhe në rastin kur kemi forma të ngjashme fjalës por që kanë tema të ndryshme për kategori të ndryshme morfologjike. Ky temëzues është renditur i pari nga 26 punimet e paraqitura në nën detyrën e temëzimit të CoNLL-18 Shared Task on Multilingual Parsing from Raw Text to Universal Dependencies (Zeman, et al., 2018).

Në këtë model gjetja e temës së fjalës është konsideruar njëlloj si problemi sekuencial i përkthimit të tekstit nga një gjuhë në tjetrën. Si hyrje e modelit të temëzuesit është një sekuencë karakteresh të fjalës së bashku me etiketat e pjesëve të ligjëratës dhe karakteristikave morfologjike dhe si dalje kemi një sekuence karakteresh që përfaqësojnë temën e fjalës. Paraqitja e hyrjes dhe daljes për fjalën rrugët janë:

INPUT: rrugët UPOS=NOUN XPOS=NNS Gender=Fem Number=Plur

OUTPUT: rrugë

Në implementimin e këtij modeli autorët kanë përdorur modelin e përkthyesit automatik OpenNMT: Open-Source Toolkit for Neural Machine Translation (Klein, et al., 2017) të implementuar në Python. Ky model është një rrjet me vëmendje të thellë enkoder-dekoder (angl. deep attentional encoder-decoder network). Në Figurën 5.2 paraqitet arkitektura e modelit enkoder-dekoder i përdorur për gjetjen e temës së fjalë. Enkoderi përdor një rrjet me shtresa dy-drejtimore LSTM, karakteret e mësuara dhe etiketat embeddings për të enkoduar karakteret e hyrjes dhe etiketat e pjesëve të ligjëratës dhe karakteristikave morfologjike në vektorë të enkoduar të të njëjtës gjatësi. Dekoderi gjeneron karakteret e sekuencës së daljes duke përdorur shtresa LSTM një-drejtimore me *input feeding attention* të aplikuar mbi daljen e enkoderit. Në *input feeding attention* peshat e mëparshme janë dhënë si hyrje në hapin tjetër të kohës për të informuar modelin për vendimet e shtrirjen se kaluar për të parandaluar në këtë mënyrë modelin të përsëris të njëjtën dalje shumë herë.

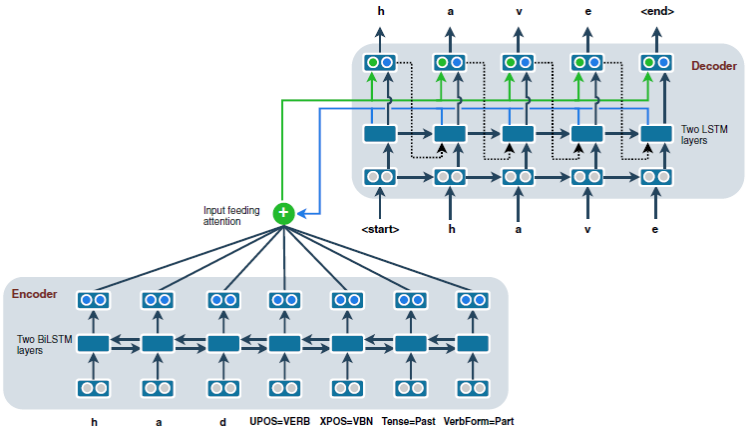


Figura 5.2 Arkitektura e modelit enkoder-dekoder (Kanerva, et al., 2020)

Për të optimizuar hiperparametrat e përdorur për trajnimin e modelit autorët kanë përdorur librarinë RBFOpt. Nga rezultatet e arritura autorët theksojnë që modeli është i qëndrueshëm dhe disa parametra kanë ndikim individual në performancën e tij. Parametrat e optimizuar që përdoren në model janë: madhësia e *embedding* 500, *dropout* 0.3, *recurrent size* 500 dhe optimizuesi Adam me *learning rate* fillestare 0.0005 dhe *learning rate decay* 0.9 pas epokës së 20 dhe madhësi të minibatch prej 64. Modeli trajnohet për 50 epoka. Gjatë trajnimit të modele ku keni një korpus të vogël trajnimi sugjerohet që madhësia e minibatch nga 64 të jetë 32 kur kemi 2000 fjali trajnimi dhe 6 për 200 fjali trajnimi.

Temëzuesi nuk shikon të gjithë fjalinë e etiketuar por vetëm një fjalë dhe pikërisht etiketat e pjesëve të ligjëratës dhe e karakteristikave morfologjike përdoret për të marrë informacionin në lidhjen me kontekstin morfologjike-sintaksore të fjalës. Pikërisht duke qenë se në hyrje ky temëzues ka nevojë për etiketat e pjesëve të ligjëratës dhe karakteristikave morfologjike është vendosur si fazë e fundit e parserit Turku Pipeline. Fillimisht mbi tekst duhet të aplikohet një parsues për gjenerimin e etiketave dhe më pas temëzuesi për gjenerimin e temën së fjalës. Kjo është një metodë e re renditje komponentësh duke qenë se sistemet e mëparshme propozojnë që temëzuesi të ekzekutohet para etiketuesit. Vendosja e temëzuesit si fazë e fundit të parserit pas etiketuesit nuk shkakton uljen e saktësisë së parserit sepse komponenti i etiketimit nuk përdor temat e fjalëve (Kanerva, et al., 2020).

KREU 6

EKSPERIMENTIMI I ALGORITMEVE DHE ANALIZA E REZULTATEVE

Në këtë kapitull prezantohet metoda e përdorur për të realizuar Opinion Mining (OM) në gjuhën shqipe dhe një etiketues morfologjik për gjuhën shqipe. Detyra e OM-së që ne do të trajtojmë është klasifikimi i opinioneve duke u bazuar në polaritetin e ndjenjës së shprehur në opinion. Etiketuesi morfologjik i propozuar realizon: segmentimin e tekstit në fjali dhe në fjalë, etiketimin e fjalëve të një fjalie me etiketën për kategorinë e pjesëve të ligjëratës që ajo fjalë i përket dhe me kategoritë gramatikore të formës së përdorur, dhe përcaktimin e temës së fjalës. Të gjitha zgjidhjet e propozuara janë vlerësuar me anë të eksperimenteve.

6.1. Klasifikimi i opinioneve në gjuhën shqipe

Të klasifikosh një opinion do të thotë të analizosh dhe të nxjerrësh polaritetin apo ngjyrimin e ndjenjës apo mendimit të shprehur nga ai. Siç kemi diskutuar në çështjen 3.3, klasifikimi i një opinioni mund të realizohet në tri nivele, në nivel dokumenti, në nivel fjalie dhe në nivel aspekti dhe entiteti. Klasat e klasifikimit mund të jenë nga më të ndryshme, si: pozitive dhe negative, pozitive, negative dhe neutrale, me pikë p.sh.: nga 1 për negative deri në 5 në pozitive, etj.

Në këtë disertacion jemi fokusuar në klasifikimin në nivel dokumenti të opinioneve në dy klasa, pozitive dhe negative. Opinionet e përdorura për klasifikim janë ruajtur në dokumente tekst. Një opinion klasifikohet si pozitiv nëse në përgjithësi ndjenja i shprehur nga ai ka një polaritet pozitiv dhe si negativ nëse në përgjithësi ndjenja e shprehur nga ai ka një polaritet negativ. Pra, klasifikimi realizohet duke realizuar një shumatore të polaritetit të shprehur nga çdo fjali me të cilin është shprehur opinioni.

6.1.1. Komponentët

Në këtë pjesë do të paraqesim në mënyrë të detajuar komponentët përbërësë të modele të përdorura për klasifikimin e opinioneve në gjuhën shqipe.

Ne kemi përdorur një model bazë për algoritme të ndryshme të të mësuarit e automatizuar (MA) dhe dy modele të tjera për dy rrjete neurale. Modeli bazë është implementuar duke përdorur algoritme MA të implementuar në platformën Weka (Witten, et al., 2017) kurse dy modelet e tjera janë implementuar duke përdorur platformat Keras dhe Tensorflow.

Në Figurën 6.1 tregohen hapat në mënyrë të përgjithshme për të ndërtuar një model klasifikimi opinionesh duke përdorur një algoritëm MA ose një rrjet neural. Si hyrje të algoritmit MA/rrjetit neural përdoret një bashkësi dokumentesh opinionesh të klasifikuara

sipas fushës së tyre dhe polaritetit të opinioni. Kjo bashkësi dokumentesh përdoret nga një algoritëm MA/rrjet neural për t'u trajnuar dhe për të gjeneruar modelin. Modeli i trajnuar më pas do të vlerësohet duke përdorur një bashkësi tjetër dokumentesh opinionesh të etiketuara.

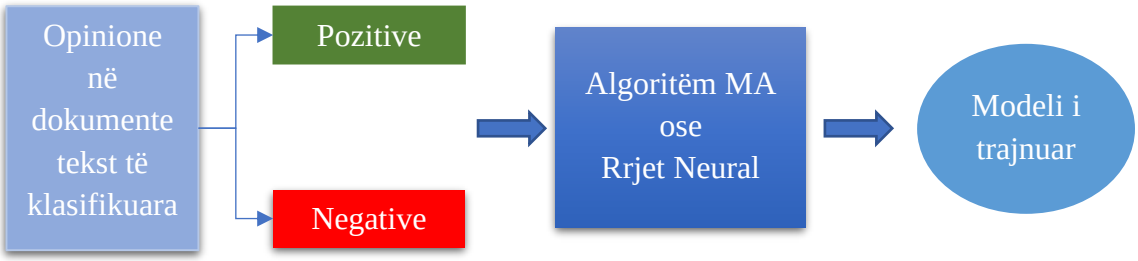


Figura 6.1 Trajnimi i një algoritmi ose rrjeti neural për të gjeneruar modelin

Në Figurën 6.2 dhe Figurën 6.3 tregohen me detaje hapat që janë ndjekur për trajnimin dhe vlerësimin e një modeli klasifikimi opinionesh duke përdorur një nga algoritmet MA dhe rrjetin neural.

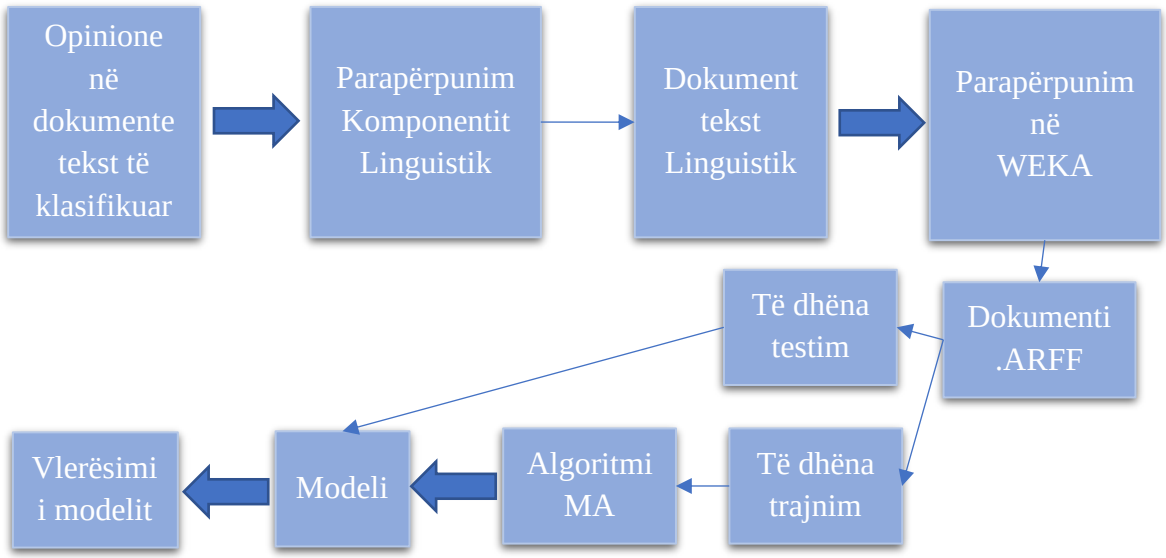


Figura 6.2 Skema e gjenerimit të modelit duke përdorur algoritmet MA

Në punime të ndryshme shkencore në fushën e OM për gjuhë të tjera, kryesisht në gjuhën angleze, referohet që përdorimi i një komponenti linguistik dhe i metodave të ndryshme të parapërpunimit rrit performancën e algoritmeve MA. Për këtë arsye dokumentet tekst të opinioneve të klasifikuara i kemi kaluar në dy faza parapërpunimi. Në fazën e parë është përdorur komponentin linguistik i shpjeguar në çështjen 6.1.3. Ky komponent linguistik bazohet në rregullat e gramatikës së gjuhës shqipe. Në fazën e dytë të parapërpunimit, dokumenti i krijuar nga komponenti linguistik përpunohet:

- Në rastin e algoritmeve MA me mjete të ndryshme parapërpunimi në platformën Weka të shpjeguara në çështjen 6.1.5.
- Në rastin e rrjeteve neurale me mjete të ndryshme parapërpunimi të shpjeguara në çështjen 6.1.7 për secilin nga rrjetet neurale.

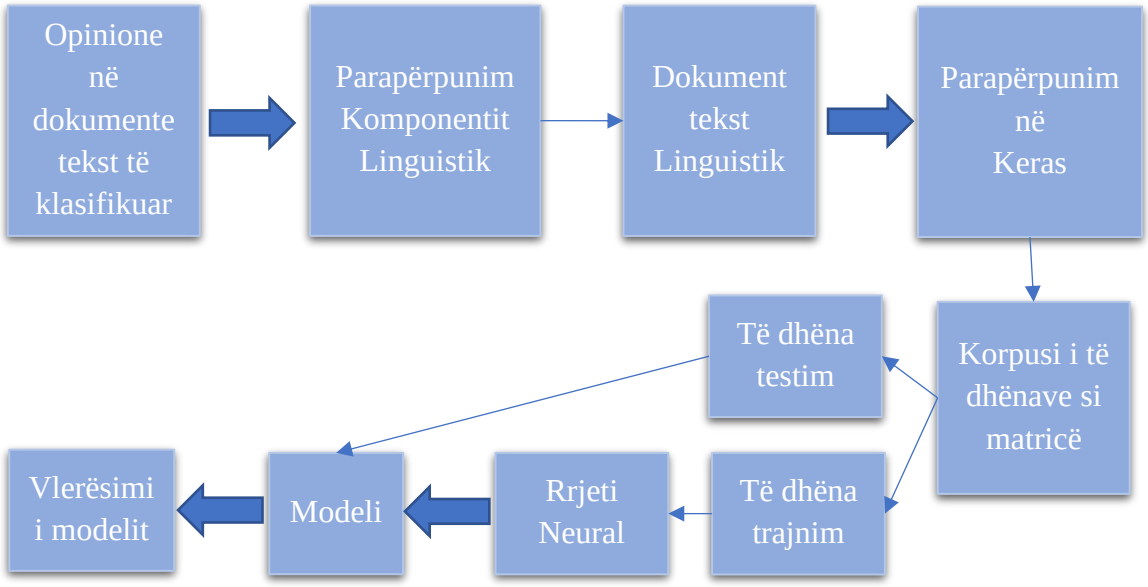


Figura 6.3 Skema e gjenerimit të modelit duke përdorur rrjetin neural

6.1.2. Korpusi i dokumenteve të opinioneve

Duke patur parasysh që në gjuhën shqipe nuk ekzistojnë korpuse të etiketuara për t'u përdorur për qëllime OM dhe veçanërisht për klasifikimin e opinioneve, ne kemi krijuar një korpus opinionesh të klasifikuara. Dokumentet e përdorura janë dokumente tekst që përmbajnë opinione rreth 5 tema diskutimesh të kohëve të fundit: politike, turizëm, biznesi i vogël, arsim, importi i mbetjeve urbane. Opinionet janë mbledhur nga artikuj në medie online të njohura në Shqipëri. Për çdo temë janë mbledhur dhe etiketuar një numër i njëjtë opinionesh të klasifikuara si pozitive dhe opinionesh të klasifikuara si negative. Një opinion është klasifikuar si pozitiv nëse polariteti në tërësi i ndjenjës së shprehur nga opinioni është pozitiv dhe si negativ nëse polariteti në tërësi i ndjenjës së shprehur nga opinioni është negativ.

Për të vlerësuar performancën e algoritmeve MA ne i kemi organizuar dhe kombinuar dokumentet e mbledhura të opinioneve në disa korpuse. Në Tabelën 6.1 tregohen në mënyrë të detajuar korpuset që janë krijuar dhe përmbajtja e secilës prej tyre.

Tabela 6.1 Korpuset e opinioneve

Kodi i korpusit	***Nr. i Temave***	***Fusha e Temave***	***Numri total i opinioneve (pozitive/negative)***
C_1	1	Turizëm	100 (50/50)
C_2	1	Arsim	100 (50/50)
C_3	1	Politikë	100 (50/50)
C_4	1	Biznesi vogël	100 (50/50)
C_5	1	Importi i mbetjeve urbane	100 (50/50)
C_6	5	Turizëm, Arsim, Politikë, Biznesi i vogël, Importi i mbetjeve urbane	500(250/250) (50 opinione pozitive dhe 50 opinione negative nga çdo temë)
C_7	5	Turizëm, Arsim, Politikë, Biznesi i vogël, Importi i mbetjeve urbane	400(200/200) (40 opinione pozitive dhe 40 opinione negative nga çdo temë)
C_8	5	Turizëm, Arsim, Politikë, Biznesi i vogël, Importi i mbetjeve urbane	300(150/150) (30 opinione pozitive dhe 30 opinione negative nga çdo temë)
C_9	5	Turizëm, Arsim, Politikë, Biznesi i vogël, Importi i mbetjeve urbane	200(100/100) (20 opinione pozitive dhe 20 opinione negative nga çdo temë)
C_10	5	Turizëm, Arsim, Politikë, Biznesi i vogël, Importi i mbetjeve urbane	100(50/50) (10 opinione pozitive dhe 10 opinione negative nga çdo temë)
C_11	4	Turizëm, Arsim, Biznesi i vogël, Importi i mbetjeve urbane	400(200/200) (50 opinione pozitive dhe 50 opinione negative nga çdo temë)
C_12	3	Turizëm, Arsim, Importi i mbetjeve urbane	300(150/150) (50 opinione pozitive dhe 50 opinione negative nga çdo temë)
C_13	3	Turizëm, Biznesi i vogël, Importi i mbetjeve urbane	300(150/150) (50 opinione pozitive dhe 50 opinione negative nga çdo temë)
C_14	2	Arsim, Politikë	200(100/100) (50 opinione pozitive dhe 50 opinione negative nga çdo temë)
C_15	2	Politikë, Biznesi i vogël	200(100/100) (50 opinione pozitive dhe 50 opinione negative nga çdo temë)
C_16	2	Turizëm, Importi i mbetjeve urbane	200(100/100) (50 pozitive dhe 50 negative nga çdo temë)
C_17	10	Turizëm, Arsim, Biznesi i vogël, Importi i mbetjeve urbane dhe 5 tema Politikë	900 (450/450) (4 temat e para: 50 pozitive dhe 50 negative, 1 temë politike: 50 pozitive dhe 50 negative, 4 tema politike: 40 pozitive dhe 40 negative)

Korpuset C_1 deri në C_5 përmbajnë opinione vetëm nga një temë e caktuar dhe janë përdorur për të analizuar performancën e algoritmeve MA për klasifikimin e opinioneve për *in-domain* OM. Korpuset C_6 deri në C_16 janë krijuar duke kombinuar tema të ndryshme dhe numër të ndryshëm opinionesh të klasifikuara si pozitive dhe negative. Këto korpuse janë përdorur për të analizuar performancën e algoritmeve MA për klasifikimin e opinioneve për *multi-domain* OM. Korpusi C_17 është korpusi që përmban numrin më të madh të opinioneve dhe është përdorur për të trajnuar dhe testuar modelet e rrjeteve neurale dhe të disa nga algoritmeve MA. Ky korpus përmban të gjitha opinionet e mbledhura.

6.1.3. Komponenti linguistik

Komponenti linguistik i përdorur bazohet në rregullat gramatikore të gjuhës shqipe dhe përbëhet nga dy faza: eliminimi i lidhëzave dhe shenjave të pikësimit dhe gjetja e rrënjë së fjalëve. Komponenti linguistik është implementuar nga Sadiku dhe Biba (2012) dhe Biba dhe Gjata (2014). Çdo dokument tekst që përmban një opinion është parapërpunuar individualisht nëpërmjet këtij komponenti linguistik dhe dokumenti i krijuar pëmban vetëm rrënjët e fjalëve. Në Figurën 6.4 tregohet skema e hapave që janë ekzekutuar nga komponenti linguistik në çdo fazë të tij.

Në gjuhën shqipe lidhëzat janë pjesë e pandryshueshme e ligjëratës dhe shërbejnë për të lidhur fjale, grupe fjalësh apo fjali midis tyre. Duke qenë se lidhëzat nuk mbartin ndonjë ngjyrim pozitiv apo negativ i kemi hequr nga dokumenti i opinionit në fazën e parapërpunimit.

Çdo dokument tekst ka kaluar në fazën e parë ku:

a. Fillimisht çdo fjalë vendoset në shkronja të vogla;
b. Janë fshirë të gjitha lidhëzat;
c. Janë fshirë karakteret speciale, si shenjat e pikësimit, dhe numrat.

Dokumenti tekst i gjeneruar nga faza e parë përdoret si hyrje në fazën e dytë të komponentit linguistik. Në fazën e dytë të komponentit linguistik, për çdo fjalë të dokumentit tekst gjendet rrënja e fjalës. Gjetja e rrënjës së fjalës ka të bëjë me segmentimin e fjalës në mënyrë automatike në njësi më të vogël kuptimore, morfema. P.sh. fjala *punoj* ndahet në morfema përbërëse: pun që është rrënja, o që është prapashtesë dhe j që është mbaresë. Algoritmi përdorur gjen rrënjën e një fjale duke eliminuar në mënyrë automatike parashtesat, prapashtesat dhe mbaresat. Ky është një algoritëm i bazuar në rregullat e fjalëformim të gjuhës shqipe. Algoritmi i implementuar nga Sadiku dhe Biba (2012) gjen rrënjën e fjalëve të thjeshta kurse algoritmi i implementuar nga Biba dhe Gjati (2014) është një zgjerim i algoritmit të parë për të gjetur rrënjën fjalëve të përbëra. Algoritmi i gjenerimit të rrënjës së fjalëve të thjeshta ka 5 hapa, secili i përbërë nga një bashkësi rregullash. Në total ky algoritëm ka të implementuar 134 rregulla për të eliminuar parashtesat, prapashtesat dhe mbaresat e fjalëve. Rregullat në algoritëm janë implementuar si kushte if-else. Fillimisht eliminohet prapashtesa ose parashtesa më e gjatë që përmban fjala duke mos e zbatuar këtë rregull nëse fjala pas eliminimit ka më pak se 2 shkronja. Në rastin kur

fjalët kanë dy zanore apo prapashtesa njëra pas tjetrës, rregulli i eliminimit të prapashtesa ose parashtesa aplikohet dy herë. Në gjuhën shqipe nuk kemi të përcaktuara rregulla të përgjithshme për gjenerimin e shumësit të fjalës dhe formave të ndryshme gjinore të fjalëve. Për këtë arsye dhe algoritmi i implementuar nuk përmban rregulla për këto dy kategori.

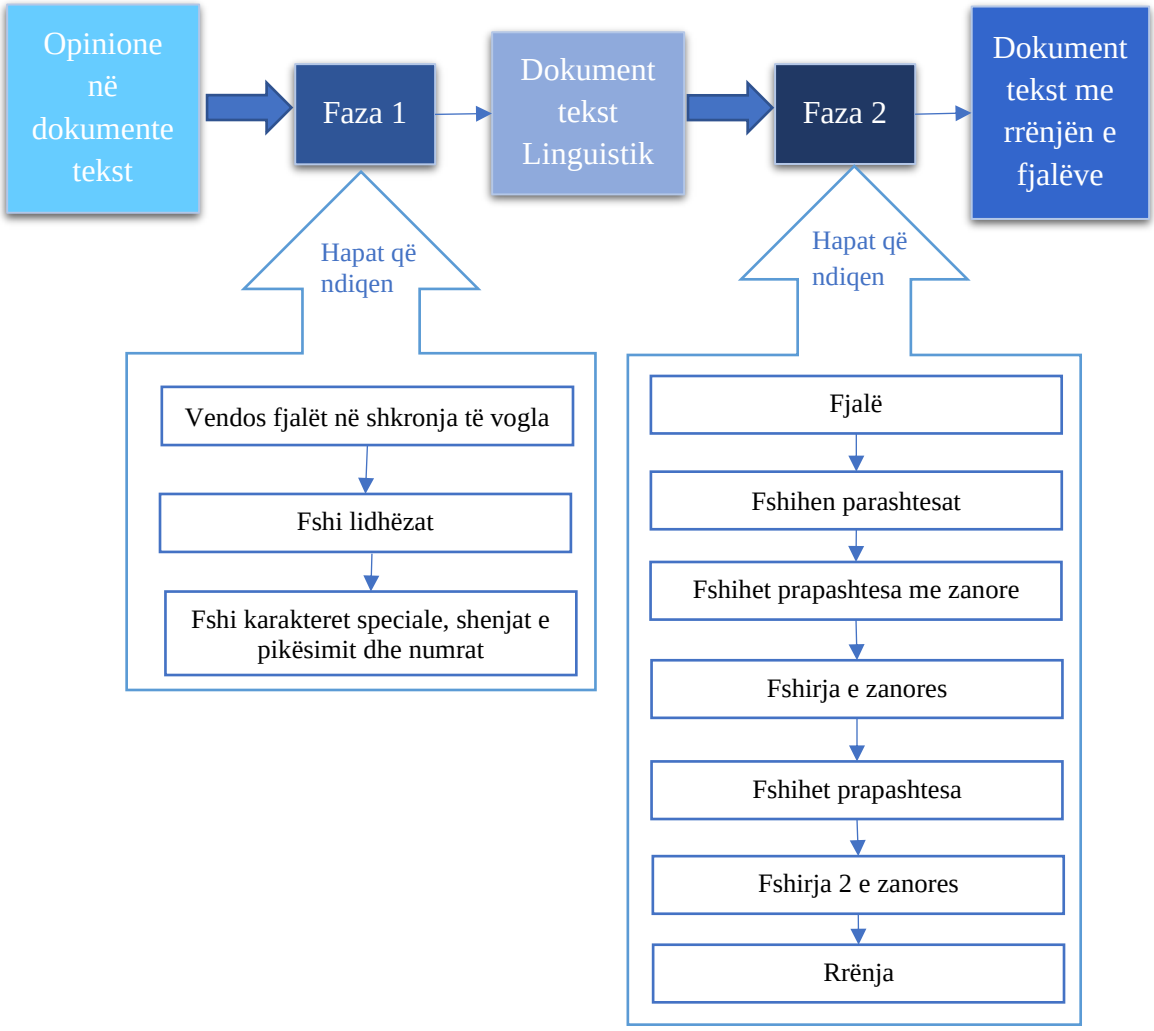


Figura 6.4 Skema e komponentit linguistik

Algoritmi për gjetjen e rrënjëve të fjalëve të përbëra është një zgjerim i algoritmit të parë dhe përfshin rregulla për gjetjen e rrënjëve të fjalëve të përbëra në gjuhën shqipe. Rregullat e implementuara janë bazuar në rregullat e fjalëformimit të fjalëve të përbëra në gjuhën shqipe. Algoritmi përmban rregulla për 3 grupe të formimit të fjalëve të përbëra në gjuhën shqipe: fjalë të formuara nga bashkimi i dy fjalëve me një vijë ndarëse, fjalë të

formuara nga një parafjalë, fjalë të formuara nga një numërues, fjalë të formuara nga bashkimi i emrave me folje, mbiemra ose emra të tjerë. Çdo rregull është implementuar si një bashkësi instruksionesh të kushtëzuara bazuar në mënyrës se si është krijuar fjala e përbërë. Në rastin kur fjala e përbërë është formuar nga bashkimi i dy fjalëve me një vijë ndarëse dhe nga bashkimi i emrave me folje, mbiemra ose emra të tjerë, fjala ndahet në dy fjalë të veçanta të thjeshta të cilave u gjendet rrënja përkatëse.

6.1.4. Algoritmet e të mësuarit e automatizuar

Detyra e klasifikimi të tekstit është një detyrë e studiuar gjerësisht në përpunimin e të dhënave, përpunimin e tekstit dhe përpunimin e gjuhës natyrale, që ka si qëllim të përcaktojë kategorinë të cilës i përket një tekst i caktuar. Në sistemet e të mësuarit e automatizuar të kontrolluar, për detyrën e klasifikimit të tekstit duhet të përcaktohet një set të dhënash trajnimi D = {X1 X2, X3, ..., Xn}, ku për çdo rekord është përcaktuar një etiketë nga një grup vlerash të ndryshme diskrete të indeksuar nga {1 ... k}. Të dhënat trajnuese përdoren për të ndërtuar një model klasifikimi, që lidh tiparet e të dhënave me një nga etiketat. Pra, për secilën instancë teksti të ri të paklasifikuar, modeli i trajnuar parashikon një etiketë (Witten, et al., 2017).

Metodat e klasifikimit kategorizohen në pesë kategori: klasifikuesit probabilistik dhe Naïve Bayes (angl. Probabilistic and Naïve Bayes Classifiers), klasifikuesit e bazuar në rregulla (angl. Rule-based Classifiers), klasifikuesit e bazuar në përafrim (angl. Proximity-based Classifiers), klasifikuesit linear (angl. Linear Classifiers) dhe klasifikuesit e pemëve vendimmarrëse (angl. Decision Tree Classifiers).

6.1.4.1. Algoritmet probabilistik dhe Naïve Bayes

Klasifikuesit probabilistik përcaktojnë për një të dhënë në hyrje një shpërndarje statistikore mbi një grup klasash, dhe jo vetëm mbi klasën më të mundshme që e dhëna në hyrje duhet t'i përkasi.

Klasifikuesi Naïve Bayes është një algoritëm i drejtpërdrejtë dhe i fuqishëm, që përdor dy grupe modelesh. Këto dy modele llogaritin probabilitetin e pasmë të një klase bazuar në shpërndarjen e fjalëve në dokumentin e tekstit. Modelet ndryshojnë nga njëri-tjetri në përdorimin ose jo të frekuencave të fjalës dhe nga veprimi i ndërmarr për të modeluar hapësirën e probabilitetit.

Klasifikuesit probabilistik dhe Naïve Bayes të marrë në konsideratë gjatë këtij punimi, të implementuar në Weka, janë: Bayesian Logistic Regression, Bayes Net, Complement Naïve Bayes, Naïve Bayes, Naïve Bayes Multinomial, Naïve Bayes Multinomial Text, Naïve Bayes Multinomial Updateable dhe Naïve Bayes Updateable.

6.1.4.2. Algoritmet e bazuar në rregulla

Klasifikuesi i bazuar në rregulla përdor një sërë rregullash për të klasifikuar të dhënën në hyrje. Ai përdor metodën nda-dhe-veço që është një metodë që bazohet në përsëritjen e procesit të gjenerimit të një rregulli për të përfaqësuar pjesët e të dhënave të trajnimit dhe më pas fshirja e këtyre të dhënave të përfaqësuara nga rregulli në korpusin e trajnimit. Ky proces përsëritet derisa nuk ka mbetur më asnjë e dhënë për t'u përfaqësuar. Këto teknika janë shumë të përdorshme, sepse janë të lehta për t'u interpretuar dhe mirëmbajtur. Një nga klasifikuesit më të përdorur i bazuar në rregulla është RIPPER i zbatuar në Weka si JRip.

Klasifikuesit e bazuar në rregulla të marrë në konsideratë gjatë këtij punimi, të implementuar në Weka janë: Conjunctive Rule, Decision Table, JRip, NNge, OneR, PART, Ridor and ZeroR.

6.1.4.3. Algoritmet e bazuar në përafrim

Klasifikuesit e bazuar në përafrim për të realizuar klasifikimin e të dhënave përdorin afërsinë e të dhënave me një nga klasat e klasifikimit. Të dhënës i përcaktohet një klasë bazuar në ngjashmërinë e metrikeve, si: produkti me pika ose metrika e kosinusit. Një e dhënë klasifikohet duke llogaritur ngjashmërinë midis kësaj të dhëne dhe të dhënave të trajnimit, dhe kësaj të dhëne i përcaktohet klasa me të cilën ka numrin më të madh të të dhënave të ngjashme.

Klasifikuesit e bazuar në afërsi të marrë në konsideratë gjatë këtij punimi, të implementuar në Weka, janë: IB1, IBK, Kstar dhe LWL.

6.1.4.4. Algoritmet linearë

Klasifikuesit linearë përcaktojnë klasën që i përket një e dhënë duke u bazuar në vlerën e kombinimit linear të karakteristikave të saj. Karakteristikat e të dhënave paraqiten në një vektor të quajtur vektori i tipareve.

Algoritmet Support Vector Machines, Regression-Based dhe Neural Network klasifikohen në këtë kategori.

Klasifikuesit linearë të marrë në konsideratë gjatë këtij punimi, të implementuar në Weka, janë: Logistic, RBF Classifier, RBF Network, SGD, SGD Text, Simple Logistic, SMO dhe Voted Perceptron.

6.1.4.5. Algoritmet e pemëve vendimmarrëse

Klasifikuesi i pemëve vendimmarrëse përdor një ndarje hierarkike të të dhënave të trajnimit duke përdorur një kusht ose një pohim mbi vlerën e atributit. Kjo ndarje përdoret për të krijuar segmentime të të dhënave të trajnimit duke u bazuar në shpërndarjen e tyre në korpus. Në të dhënat tekst këto kushte ose pohime lidhen me mungesën ose praninë e një ose më shumë fjalëve në dokument.

Klasifikuesit e pemëve të vendimmarrëse të marrë në konsideratë gjatë këtij punimi, të implementuar në Weka, janë: Decision Stump, HoeffdingTree, J48, LMT, Random Forest, Random Tree, REPTree and SimpleCart.

Gjithashtu janë marrë në konsideratë dhe disa algoritme të implementuar në dy klasa të tjera në Weka: klasa Meta dhe klasa Misc. Klasifikuesit në klasën Meta janë kombinime të klasifikuesve të ndryshëm. Klasifikuesit e klasës Meta të marrë në konsideratë janë: AdaBoost M1, Attribute Selected Classifier, Bagging, Classification Via Regression, Filtered Classifier, Interative Classifier Optimizer, Logit Boster, Multi Class Classifier, Multi Class Classifier Updateable, Random Committee, Randomizable Filtered Classifier, Random Sub Space, Real Ada Boost dhe Vote. Klasifikuesit në klasën Misc të marrë në konsideratë janë: FLR, Hyper Pipes dhe Input Mapped Classifier classifiers.

6.1.5. Parapërpunimi në Weka

Dokumentet tekst të secilit korpus të parapërpunuar me komponentin linguistik ngarkohen në Weka dhe mbi to aplikohet një fazë e dytë parapërpunimi për të krijuar një dokument me prapashtesë .arff. Dokumenti me prapashtesën .arff përdoret për trajnimin dhe vlerësimin e modeleve të algoritmeve MA.

Dokumentet tekst të secili korpus janë ngarkuar në Weka duke përdorur klasën *textDirectoryLoader*. Kjo klasë ngarkon të gjitha dokumentet tekst në një direktori duke përdorur emrat e nën-direktorive si etiketa të klasave dhe e ruan përmbajtjen në një stringë atributesh.

Më pas është aplikuar filteri *StringToWordVector*, që konverton të gjithë stringat e atributeve në një vektor fjalësh, që paraqet përsëritjen e një fjale në tekst. Filtri na mundëson të zgjedhim karakteristika të ndryshme për t'i aplikuar mbi tekst. Për të identifikuar ndikimin e karakteristikave të ndryshme dhe për të përcaktuar modelin me saktësi më të lartë, ne kemi përdorur 6 konfigurime të karakteristikave në këtë filtër si më poshtë:

a. WordTokenizer;
b. WordTokenizer dhe TF-IDF;
c. n-gram me vlerat min=1 dhe max=2;
d. TF-IDF dhe n-gram me vlerat min=1 dhe max=2;
e. n-gram me vlerat min=1 dhe max=3;
f. TF-IDF dhe n-gram me vlerat min=1 dhe max=3.

Pas aplikimit të këtij filtri është krijuar dokumenti me prapashtesë .arff i cili përdoret për të trajnuar një model duke përdorur një nga algoritmet MA.

6.1.6. Vlerësimi eksperimental i algoritmeve të të mësuarit e automatizuar

Në këtë pjesë prezantohen rezultatet e eksperimenteve të realizuara dhe analizohen rezultatet e performancës së algoritmeve MA të zgjedhur për klasifikimin e opinioneve. Ne kemi përzgjedhur 50 algoritme MA për t'iu vlerësuar performancën në klasifikimin e opinioneve në gjuhën shqipe si opinione pozitive dhe negative.

Për të trajnuar dhe vlerësuar një model në Weka është përdorur Experimenter-i me 10-folds Cross-validation. 10-folds Cross-validation e ndan dokumentin .arff në 10 grupe, ku për çdo vlerësim 9 grupe përdoren për trajnim dhe 1 grup përdoret për vlerësim të modelit të trajnuar.

6.1.6.1. Vlerësimi i algoritmeve për in-domain Opinion Mining

Qëllimi i këtij grupi eksperimentesh është të vlerësohet performanca e 50 algoritmeve të përzgjedhur në korpuse që përmbajnë opinione vetëm nga një temë, ajo që në gjuhën angleze njihet si *in-domain* OM. Janë trajnuar dhe vlerësuar nga një model për secilin nga 50 algoritmet e zgjedhura duke përdorur korpuset C_1 deri në C_5. Korpusi i opinioneve në këto eksperimente është parapërpunuar duke përdorur të dyja fazat e komponentit linguistik dhe konfigurimin a) të shpjeguar në çështjen 6.1.5 për filtrin *StringToWordVector* në Weka.

Në Tabelën 6.2 tregohen rezultatet e eksperimenteve për çdo algoritëm për çdo korpus në term të përqindjes së instancave të klasifikuara në mënyrë korrekte. Për çdo korpus kemi theksuar me ngjyrë të zezë rezultatin e algoritmit që ka vlerën më të lartë të përqindjes së instancave të klasifikuara në mënyrë korrekte për atë korpus. Duke analizuar rezultatet e eksperimenteve, modele të algoritmeve të ndryshme kanë performancën më të mirë për korpuse të ndryshme. Dhe përkatësisht:

- Për korpusin C_1, dy algoritmet, Logistic dhe Multi Class Classifier, kanë rezultatin më të mirë të përqindjes së instancave të klasifikuara në mënyrë korrekte prej 94%;
- Për korpusin C_2, algoritmi Hyper Pipes ka rezultatin më të mirë të përqindjes së instancave të klasifikuara në mënyrë korrekte prej 92%;
- Për korpusin C_3, algoritmi RBF Classifier ka rezultatin më të mirë të përqindjes së instancave të klasifikuara në mënyrë korrekte prej 79%;
- Për korpusin C_4 dhe C_5, algoritmi RBF Network ka rezultatin më të mirë të përqindjes së instancave të klasifikuara në mënyrë korrekte, por me vlera të ndryshme përkatësisht prej 86% për korpusin C_4 dhe 89% për korpusin C_5;

Tabela 6.2 Rezultatet në term të përqindjen së instancave të klasifikuara në mënyrë korrekte për in-domain Opinion Mining

Algoritmi	***C_1***	***C_2***	***C_3***	***C_4***	***C_5***
BayesianLogisticRegression	89	85	77	74	78
BayesNet	87	59	68	49	71
ComplementNaiveBayes	87	89	77	79	85
NaiveBayes	86	83	71	74	78
NaiveBayesMultinomial	87	89	77	79	85
NaiveBayesMultinomialText	50	50	50	50	50
NaiveBayesMultinomialUpdateable	87	89	77	79	85
NaiveBayeslUpdateable	86	83	71	74	78
Logistic	**94**	84	66	82	79
RBFClassifier	85	77	**79**	67	75
RBFNetwork	49	88	66	**86**	**89**
SGD	88	84	75	75	81
SimpleLogistic	84	74	66	63	71
SMO	88	82	74	75	78
VotedPerceptron	86	76	74	71	75
IB1	59	65	62	60	62
IBK	60	65	62	60	62
KStar	58	65	62	60	62
LWL	69	50	75	58	57
AdaBoostM1	80	71	70	52	76
AttributeSelectedClassifier	84	63	66	53	71
Bagging	83	62	75	62	66
ClassificationViaRegression	78	54	71	54	64
FilteredClassifier	85	56	68	49	60
InterativeClassifierOptimizer	79	51	75	50	71
LogitBoster	81	67	61	62	76
MultiClassClassifier	**94**	84	66	82	79
MultiClassClassifierUpdateable	88	84	75	75	81
RandomCommittee	77	71	69	67	69
RandomizableFilteredClassifier	55	60	58	55	49
RandomSubSpace	82	65	73	64	64
RealAdaBoost	78	70	67	58	68
FLR	89	91	66	85	81
HyperPipes	92	**92**	67	85	80
ConjunctiveRule	62	50	61	59	58
DecisionTable	69	54	71	57	61
JRip	76	53	70	64	70
NNge	85	82	69	64	65
OneR	69	49	75	56	65
PART	75	67	63	58	65
Ridor	75	65	57	52	64
ZeroR	50	50	50	50	50
DecisionStump	61	50	75	57	56
HoeffdingTree	86	81	74	73	82
J48	87	67	58	60	62
LMT	84	74	66	64	70
RandomForest	86	76	76	65	74
RandomTree	57	64	59	64	67
REPTree	74	51	71	54	54
SimpleCart	73	53	75	54	62

Kemi 5 algoritme që kanë performancë më të mirë, që varion nga 79% në 94%. Një fakt interesant, që rrjedh nga analizimi i rezultateve, është që rezultati më i mirë dhe rezultati më i keq i secilit algoritëm në 5 korpuset ka një diferencë të konsiderueshme. Kjo diferencë për secilin nga pesë algoritmet më performant është:

- Logistic -> 28%;
- Multi Class Classifier -> 28%;
- Hyper Pipes -> 25%;
- RBF Classifier -> 12%;
- RBF Network -> 40%.

Algoritmet Logistic dhe Multi Class Classifier kanë të njëjtën performancë. Algoritmi RBF Classifier ka performancën më të mirë në rastin e korpusit C_1, prej 85%, por ka performancë më të ulët, prej 79% për korpusin C_3 ku është dhe algoritmi më i mirë për këtë korpus.

Për të renditur algoritmet nga më performanti tek më pak performanti, për çdo korpus ne kemi vlerësuar algoritmet me pikë nga 5 në 1, duke u bazuar në rezultatet e tyre të përqindjes së instancave të klasifikuara në mënyrë korrekte. Ky vlerësim është përdorur për të llogaritur një vlerë mesatare të ponderuar të përqindjes së instancave të klasifikuara në mënyrë korrekte për çdo algoritëm. Për shembull, nëse marrim në konsideratë rezultatet e algoritmeve për korpusin C_1, ne kemi renditur algoritmet nga më performanti në atë më pak performant. Në këtë rast algoritmet më performant janë Logistic dhe Multi Class Classifier me 94% instanca të klasifikuara në mënyrë korrekte, dhe i kemi vlerësuar me 5 pikë. Hyper Pipes është algoritmi pasardhës më pak performant me 92% instanca të klasifikuara në mënyrë korrekte, dhe e kemi vlerësuar me 4 pikë. Më pas kemi algoritmin RBF Classifier me 85% instanca të klasifikuara në mënyrë korrekte, dhe e kemi vlerësuar me 3 pikë. Dhe si algoritëm më pak performant është RBF Network me 49% instanca të klasifikuara në mënyrë korrekte, dhe e kemi vlerësuar me 2 pikë. Kjo skemë është ndjekur edhe për rezultatet e algoritmeve në korpuset e tjera. Llogaritja e vlerës mesatare të ponderuar të përqindjes së instancave të klasifikuara në mënyrë korrekte për çdo algoritëm është paraqitur në Tabelën 6.3 .

Tabela 6.3 Renditja e algoritmeve më performant

Algoritmi	***C_1***	***C_2***	***C_3***	***C_4***	***C_5***	***Mesatarja e ponderuar***
Hyper Pipes	92*4	92*5	67*4	85*4	80*4	83.62
Logistic	94*5	84*3	66*3	79*3	79*3	82.53
Multi Class Classifier	94*5	84*3	66*3	79*3	79*3	82.53
RBF Network	49*2	88*4	66*3	89*5	89*5	80.16
RBF Classifier	85*3	77*2	79*5	75*2	75*2	77.71

Në Tabelën 6.3 algoritmet janë paraqitur në mënyrë të renditur në rendin zbritës të vlerës mesatare të ponderuar të përqindjes së instancave të klasifikuara në mënyrë korrekte. Duke u bazuar në këtë skemë llogaritje, algoritmi Hyper Pipes është algoritmi me performancën më të mirë, i ndjekur nga Logistic dhe Multi Class Classifier që kanë të njëjtin rezultat, RBF Network dhe RBF Classifier.

Nga rezultatet e eksperimenteve, nuk mund të përcaktojmë një algoritëm të vetëm që performon më mirë. Për të identifikuar ndonjë ndryshim statistikor në performancë midis këtyre algoritmeve kemi realizuar një eksperiment të vlerësimit të kryqëzuar në Weka. Për të realizuar këtë eksperiment, algoritmi Hyper Pipes, si algoritmi me vlerë të mesatares së ponderuar të përqindjes së instancave të klasifikuara në mënyrë korrekte më të lartë, është përcaktuar si algoritëm bazë. Ky eksperiment është realizuar duke përdorur *Weka Experimenter tool*, 5 algoritmet më performant, 10 *cross-validation* dhe 10 *repetitions*.

Në Tabelën 6.4 tregohen rezultatet e eksperimentit të realizuar. Rezultatet përcaktojnë që për korpuset C_1 dhe C_2 nuk ka ndonjë diferencë në performancën e këtyre algoritmeve. Në rezultatin e algoritmit RBF Classifier në korpusin C_3 vërehet një "v", që do të thotë që ky algoritëm në këtë korpus performon më mirë se algoritmi bazë Hyper Pipes. Gjithashtu edhe në rezultatin e algoritmit RBF Network për korpusin C_5 vërehet e njëjta shenjë, që do të thotë që edhe ky algoritëm në këtë korpus performon më mirë se algoritmi bazë Hyper Pipes. Në rezultatin e algoritmit RBF Classifier në korpusin C_4 vërehet një "*", që do të thotë që ky algoritëm ka perfomancë më të ulët në këtë korpus se algoritmi bazë Hyper Pipes.

Si përfundim, mund të themi që edhe me këtë eksperiment nuk mund të përcaktojmë vetëm një algoritëm që performon më mirë.

Tabela 6.4 Rezultatet e eksperimentit për vlerësimin e kryqëzuar

Algoritmi	***C_1***	***C_2***	***C_3***	***C_4***	***C_5***
Hyper Pipes	89.90	90.40	64.40	83.90	80.90
Logistic	92.20	84.60	66.70	81.10	83.70
Multi Class Classifier	92.20	84.60	66.70	81.10	83.70
RBF Network	90.50	88.80	63.90	83.80	**89.70 v**
RBF Classifier	85.80	83.50	**76.60 v**	**69.70 ***	75.80

6.1.6.2. Vlerësimi i algoritmeve për multi-domain Opinion Mining

Qëllimi i këtij grupi eksperimentesh është të vlerësohet performanca e 50 algoritmeve të përzgjedhur në korpuse që përmbajnë opinione nga tema të ndryshme dhe me numër të ndryshëm opinionesh. Janë trajnuar dhe vlerësuar nga një model për secilin nga 50 algoritmet e zgjedhura duke përdorur korpuset C_6 deri në C_16. Në realizimin e këtyre eksperimenteve janë përdorur të njëjtat parametra si në grupin e parë të eksperimenteve. Korpusi i opinioneve në këto eksperimente është parapërpunuar duke përdorur të dyja fazat

e komponentit linguistik dhe konfigurimin a) të shpjeguar në çështjen 6.1.5 për filtrin *StringToWordVector* në Weka.

Në Tabelën 6.5 dhe Tabelën 6.6 paraqiten rezultatet e eksperimenteve për çdo algoritëm të përzgjedhur për çdo korpus në term të përqindjes së instancave të klasifikuara në mënyrë korrekte. Për çdo korpus kemi theksuar me ngjyrë të zezë rezultatin e algoritmit që ka vlerën më të lartë të përqindjes së instancave të klasifikuara në mënyrë korrekte për atë korpus. Nga analizimi i rezultateve, edhe në këtë rast nuk kemi vetëm një algoritëm që performon më mirë, por për korpuse të ndryshme kemi algoritme të ndryshme që kanë vlerën më të lartë të përqindjes së instancave të klasifikuara në mënyrë korrekte. Dhe përkatësisht:

- Për korpusin C_6, C_7, C_8, C_9, C_11, C_12 dhe C_13, tri algoritmet: Complement Naïve Bayes, Naïve Bayes Multinomial dhe Naïve Bayes Multinomial Updateable, kanë rezultatin më të mirë të përqindjes së instancave të klasifikuara në mënyrë korrekte, por në vlera të ndryshme për korpuse të ndryshme;
- Për korpusin C_10, algoritmi RBF Network ka rezultatin më të mirë të përqindjes së instancave të klasifikuara në mënyrë korrekte prej 79%;
- Për korpusin C_14, algoritmi Logistic ka rezultatin më të mirë të përqindjes së instancave të klasifikuara në mënyrë korrekte prej 80.50%;
- Për korpusin C_15, algoritmi Hyper Pipes ka rezultatin më të mirë të përqindjes së instancave të klasifikuara në mënyrë korrekte prej 78%;
- Për korpusin C_16, algoritmi SGD ka rezultatin më të mirë të përqindjes së instancave të klasifikuara në mënyrë korrekte prej 84.50%.

Kemi 7 algoritme të cilët kanë performancën më të mirë, që varion nga 78% në 85.7% instanca të klasifikuara në mënyrë korrekte. Algoritmet Naïve Bayes kanë të njëjtën performancë. Edhe në këtë rast, fakt interesant që rrjedh nga analizimi i rezultateve është që rezultati më i mirë dhe rezultati më i keq i secilit nga 7 algoritmet më performant në 11 korpuset ka një diferencë të konsiderueshme. Kjo diferencë për secilin algoritëm është:

- Complement Naïve Bayes, Naïve Bayes Multinomial and Naïve Bayes Multinomial Updateable -> 15.2%;
- RBF Network -> 32%
- Logistic -> 14.1%
- Hyper Pipes -> 22.2%; c
- SGD -> 17.5 %.

Për të renditur këto 7 algoritme nga më performanti në më pak performantin për cdo korpus, kemi realizuar vlerësimin me pikë dhe kemi llogaritur vlerën mesatare të ponderuar të përqindjes së instancave të klasifikuara në mënyrë korrekte. Skema e përdorur për këtë llogaritje është e njëjta që u përdor në eksperimentet e trajtuara në çështjen 6.1.6.1. Në Tabelën 6.7 paraqitet vlera mesatare e ponderuar e përqindjes së instancave të klasifikuara në mënyrë korrekte për çdo algoritëm dhe renditja e algoritmeve në rend zbritës për këtë vlerë.

Tabela 6.5 Rezultatet në term të përqindjes së instancave të klasifikuara në mënyrë korrekte për Multi-domain Opinion Mining

Algoritmi	***C_6***	***C_7***	***C_8***	***C_9***	***C_10***	***C_11***
BayesianLogisticRegretion	77.0	77.5	77.7	78.0	73.0	78.0
BayesNet	66.0	64.0	54.3	53.5	51.0	61.8
ComplementNaiveBayes	**78.6**	**80.3**	**81.7**	**80.0**	75.0	**83.0**
NaiveBayes	73.0	75.8	75.3	72.5	72.0	78.8
NaiveBayesMultinomial	**78.6**	**80.3**	**81.7**	**80.0**	75.0	**83.0**
NaiveBayesMultinomialText	50.0	50.0	50.0	50.0	50.0	50.0
NaiveBayesMultinomialUpdateab	**78.6**	**80.3**	**81.7**	**80.0**	75.0	**83.0**
NaiveBayeslUpdateable	73.0	75.8	75.3	72.5	72.0	78.8
Logistic	66.6	78.8	76.7	77.5	78.0	77.0
RBFClassifier	73.6	74.5	74.3	71.0	59.0	78.5
RBFNetwork	73.4	74.8	76.7	78.0	**79.0**	51.8
SGD	77.4	77.5	77.0	75.0	66.0	77.3
SimpleLogistic	66.4	70.0	68.7	58.5	52.0	68.0
SMO	76.0	75.0	76.0	73.5	66.0	77.0
VotedPerceptron	69.2	74.0	72.3	68.0	56.0	73.3
IB1	61.8	62.5	63.7	61.0	59.0	59.3
IBK	60.4	61.5	64.3	61.0	59.0	59.8
Kstar	61.8	62.3	64.3	62.0	57.0	61.0
LWL	55.6	54.0	52.0	53.0	55.0	57.8
AdaBoostM1	60.6	62.8	61.3	53.5	56.0	60.0
AttributeSelectedClassifier	60.6	64.0	63.3	57.0	63.0	64.0
Bagging	65.4	69.0	63.7	64.5	56.0	70.8
ClassificationViaRegression	60.2	58.0	51.7	51.5	50.0	60.3
FilteredClassifier	62.6	62.5	54.3	53.5	51.0	61.0
InterativeClassifierOptimizer	63.6	68.8	59.7	59.5	52.0	65.8
LogitBoster	66.0	71.0	63.7	61.0	51.0	69.3
MultiClassClassifier	66.6	78.8	76.7	77.5	78.0	77.0
MultiClassClassifierUpdateable	77.4	77.5	77.0	75.0	66.0	77.3
RandomCommittee	67.0	71.5	68.7	67.5	59.0	67.8
RandomizableFilteredClassifier	51.8	54.0	54.0	56.5	56.0	54.8
RandomSubSpace	72.0	66.5	66.0	63.0	59.0	71.5
RealAdaBoost	65.4	68.3	62.0	61.0	56.0	66.0
FLR	57.8	62.0	67.7	78.0	75.0	65.0
HyperPipes	55.8	62.3	64.7	78.5	78.0	63.5
ConjunctiveRule	54.4	53.0	52.0	51.5	54.0	54.8
DecisionTable	63.4	60.5	59.7	55.0	55.0	63.8
Jrip	58.4	57.5	53.3	48.0	47.0	62.3
Nnge	61.0	67.5	67.0	70.5	67.0	64.5
OneR	58.0	57.3	59.0	42.5	45.0	57.3
PART	65.6	61.3	61.3	62.0	51.0	62.5
Ridor	56.6	60.0	55.7	55.5	45.0	60.0
ZeroR	50.0	50.0	50.0	50.0	50.0	50.0
DecisionStump	52.4	52.5	52.7	49.0	55.0	56.5
HoeffdingTree	72.8	73.8	79.3	79.5	71.0	76.5
J48	61.2	61.5	64.7	65.5	59.0	63.3
LMT	68.0	70.0	67.3	58.5	55.0	68.8
Random Forest	74.2	75.5	76.7	72.0	66.0	78.0
RandomTree	57.8	56.3	55.7	60.0	50.0	58.0
REPTree	60.6	58.5	57.7	58.5	56.0	57.8
SimpleCart	61.0	62.5	61.0	57.5	56.0	63.3

Tabela 6.6 Rezultatet në term të përqindjes së instancave të klasifikuara në mënyrë korrekte për Multi-domain Opinion Mining

Algoritmi	***C_12***	***C_13***	***C_14***	***C_15***	***C_16***
BayesianLogisticRegretion	81.7	77.3	77.0	71.5	82.5
BayesNet	68.7	60.3	57.0	56.5	79.0
ComplementNaiveBayes	**85.7**	**81.7**	75.0	70.5	81.5
NaiveBayes	80.3	76.7	72.0	69.5	79.0
NaiveBayesMultinomial	**85.7**	**81.7**	75.0	70.5	81.5
NaiveBayesMultinomialText	50.0	50.0	50.0	50.0	50.0
NaiveBayesMultinomialUpdateab	**85.7**	**81.7**	75.0	70.5	81.5
NaiveBayeslUpdateable	80.3	76.7	72.0	69.5	79.0
Logistic	82.3	79.3	**80.5**	73.0	84.0
RBFClassifier	77.7	76.0	72.5	65.5	81.5
RBFNetwork	49.3	50.7	49.5	47.0	54.0
SGD	80.0	76.3	75.5	71.0	**84.5**
SimpleLogistic	74.7	65.3	65.0	55.5	74.5
SMO	78.0	76.7	74.5	68.0	81.0
VotedPerceptron	81.0	71.0	65.5	69.0	77.0
IB1	59.7	63.7	61.0	61.5	61.0
IBK	59.7	64.3	60.5	60.5	60.5
Kstar	60.0	64.0	60.0	61.5	60.5
LWL	60.7	57.7	58.0	60.0	65.5
AdaBoostM1	69.7	61.3	56.5	52.0	78.5
AttributeSelectedClassifier	76.3	61.7	64.5	55.0	75.0
Bagging	71.0	72.3	65.0	59.0	76.5
ClassificationViaRegression	72.7	65.3	49.5	59.0	72.0
FilteredClassifier	71.0	61.3	58.0	56.5	77.0
InterativeClassifierOptimizer	74.0	65.0	52.5	56.0	73.5
LogitBoster	75.0	67.0	54.5	59.0	74.5
MultiClassClassifier	82.3	79.3	80.5	73.0	84.0
MultiClassClassifierUpdateable	80.0	76.3	75.5	71.0	84.5
RandomCommittee	74.0	63.0	70.0	66.5	71.5
RandomizableFilteredClassifier	53.0	54.0	47.5	53.0	59.5
RandomSubSpace	75.0	72.0	68.5	62.5	77.0
RealAdaBoost	71.0	65.7	60.0	61.5	79.5
FLR	72.3	73.7	78.5	75.0	79.5
HyperPipes	71.3	66.7	78.5	**78.0**	74.0
ConjunctiveRule	54.7	59.7	49.5	50.5	59.5
DecisionTable	71.0	59.7	64.5	55.0	72.5
Jrip	66.7	63.7	62.5	52.0	77.0
Nnge	67.0	67.3	77.0	65.0	71.0
OneR	63.7	56.3	58.0	57.5	64.5
PART	72.0	63.3	63.0	60.0	65.5
Ridor	62.7	60.0	53.0	50.0	69.0
ZeroR	50.0	50.0	50.0	50.0	50.0
DecisionStump	57.3	56.7	54.5	62.0	66.0
HoeffdingTree	81.0	74.7	78.5	72.0	71.5
J48	68.3	61.3	64.0	60.0	69.0
LMT	76.3	65.3	65.5	55.5	74.5
Random Forest	80.7	78.3	75.0	67.0	82.0
RandomTree	63.7	60.0	59.5	54.5	51.5
REPTree	67.0	64.0	56.5	54.0	69.0
SimpleCart	69.7	71.7	55.5	58.5	76.0

Tabela 6.7 Renditja e algoritmeve më performantë

Algorimi	***C_6***	***C_7***	***C_8***	***C_9***	***C_10***	***C_11***	***C_12***	***C_13***	***C_14***	***C_15***	***C_16***	**Mesatarja e ponderuar**
ComplementNaiveBayes	78.6*5	80.3*5	81.7*5	80*5	75*3	83*5	85.7*5	81.7*5	75*2	70.5*2	81.5*3	80.33
NaiveBayesMultinomial	78.6*5	80.3*5	81.7*5	80*5	75*3	83*5	85.7*5	81.7*5	75*2	70.5*2	81.5*3	80.33
NaiveBayesMultinomialUpdateable	78.6*5	80.3*5	81.7*5	80*5	75*3	83*5	85.7*5	81.7*5	75*2	70.5*2	81.5*3	80.33
Logistic	66.6*2	78.6*4	76.7*3	77.5*2	78*4	77*3	82.3*4	79.3*4	80.5*5	73*4	84*4	78.29
SGD	77.4*4	77.5*3	77*4	75*1	66*2	77.3*4	80*3	76.3*3	75.5*3	71*3	84.5*5	77.06
HyperPipes	55.8*1	62.3*1	64.7*1	78.5*4	78*4	63.5*2	71.3*2	66.7*2	78.5*4	78*5	74*2	73.70
RBFNetwork	73.4*3	74.8*2	76.7*2	78*3	79*5	51.8*1	49.3*1	50.7*1	49.5*1	47*1	54*1	69.25

Tabela 6.8 Rezultatet e eksperimentit për vlerësimin e kryqëzuar

Algoritmi	***C_6***	***C_7***	***C_8***	***C_9***	***C_10***	***C_11***	***C_12***	***C_13***	***C_14***	***C_15***	***C_16***
ComplementNaiveBayes	76.86	79.32	81.13	79.35	76.4	83.00	85.97	79.4	75.25	70.50	82.70
NaiveBayesMultinomial	76.86	79.32	81.13	79.35	76.4	83.00	85.97	79.4	75.25	70.50	82.70
NaiveBayesMultinomialUpdate	76.86	79.32	81.13	79.35	76.4	83.00	85.97	79.4	75.25	70.50	82.70
Logistic	**64.3***	74.20	76.03	75.45	79.5	79.17	82.90	77.9	78.80	73.90	81.30
SGD	76.86	78.17	76.93	75.75	68.2	**78.20***	81.57	76.83	76.30	71.30	82.90
HyperPipes	**55.20***	**62.70***	**64.80***	78.15	82.00	**63.50***	**70.70***	**66.60***	78.20	76.35	**74.60***
RBFNetwork	72.06	74.65	**74.50***	78.65	82.90	**74.60***	**77.60***	**73.10***	**83.20v**	**78.30v**	78.75

Algoritmet Complement Naïve Bayes, Naïve Bayes Multinomial dhe Naïve Bayes Multinomial Updateable kanë rezultatin më të lartë prej 80.33%, të ndjekurë prej Logistic me një diferencë prej 2.04%, nga SGD dhe Hyper Pipes, dhe algoritmi RBF Network i renditur i fundit.

Edhe në këtë rast kemi realizuar eksperimentin e vlerësimit të kryqëzuar për të identifikuar ndonjë ndryshim statistikor në performancën e këtyre algoritmeve. Si algoritëm bazë është zgjedhur algoritmi Complement Naïve Bayes. Rezultatet e eksperimenti janë paraqitur në Tabelën 6.8.

Për korpusin C_9 dhe C_10 algoritmet nuk kanë diferencë në performancë. Në rezultatet e algoritmit Hyper Pipes për korpuset C_6, C_7, C_8, C_11, C_12, C_13 dhe C_16 vërehet një "*", që do të thotë që ky algoritëm në këto korpuse ka performancë më të keqe se algoritmi bazë Complement Naïve Bayes. Gjithashtu, "*" paraqitet dhe në rezultatin e algoritmit Logistic për korpusin C_6, në rezultatet e algoritmit RBF Network për korpuset C_8, C_11, C_12 dhe C_13 dhe në rezultatin e algoritmit SGD për korpusin C_11. Në këto raste këto algoritme kanë performancë më të ulët se algoritmi bazë Complement Naïve Bayes. Në rezultatet e algoritmit RBF Network për korpuset C_14 dhe C_15 kemi një "v", që do të thotë që ky algoritëm në këtë rast ka performancë më të mirë se algoritmi bazë Complement Naïve Bayes. Edhe me anë të këtij eksperimenti nuk mund të përcaktojmë një algoritëm të vetëm që performon më mirë.

6.1.6.3. Vlerësimi i karakteristikave të parapërpunimit në performancën e algoritmeve

Në Tabelën 6.9 është paraqitur për çdo algoritëm vlera mesatare e eksperimenteve për *in-domain* OM, vlera mesatare eksperimenteve *multi-domanin* OM dhe vlera mesatare totale për të dy grupet e eksperimenteve në term të përqindjes së instancave të klasifikuara në mënyrë korrekte. Në tabelë algoritmet janë renditur në rendin zbritës sipas vlerës së mesatares totale.

Më tej kemi realizuar një grup të tretë eksperimentesh për të investiguar më shumë në rolin që kanë karakteristika të ndryshme parapërpunimi në performancën e algoritmeve. Për të realizuar këtë grup eksperimentesh kemi marrë në konsideratë algoritmet më performantë të dy grupeve të para të eksperimenteve: Naïve Bayes Multinomial, Logistic, SGD, Hyper Pipes, dhe RBF Network dhe shtatë algoritme të tjerë: Bayesian Logistic Regression, SMO, Random Forest, Voted Perceptron, Simple Lo-gistic, J48, dhe IBK, të cilët kanë rezultate të mira në eksperimentet e realizuara dhe janë shumë popullor në punë të ndryshme kërkimore në analizimin e ndjenjave. Algoritmi Multi-Class Classifier nuk është marrë parasysh në këto eksperimente sepse ai përdor algoritmin Logistic. Eksperimentet janë realizuar duke përdorur korpuset: C_1, C_2, C_3, C_4, C_5, C_6 dhe C_17.

Tabela 6.9 Vlera mesatare e rezultateve eksperimentale

Algoritmi	***Mesatare 1***	***Mesatare 2***	***Mesatare totale***
ComplementNaiveBayes	83.40	79.35	80.62
NaiveBayesMultinomial	83.40	79.35	80.62
NaiveBayesMultinomialUpdateable	83.40	79.35	80.62
Logistic	81.00	77.61	78.67
MultiClassClassifier	81.00	77.61	78.67
BayesianLogisticRegretion	80.60	77.38	78.39
SGD	80.60	76.13	77.53
MultiClassClassifierUpdateable	80.60	76.13	77.53
HoeffdingTree	79.20	75.50	76.66
SMO	79.40	74.70	76.17
NaiveBayes	78.40	74.98	76.05
NaiveBayeslUpdateable	78.40	74.98	76.05
Random Forest	75.40	75.03	75.15
FLR	82.40	71.32	74.78
HyperPipes	83.20	70.11	74.20
RBFClassifier	76.60	73.10	74.19
VotedPerceptron	76.40	70.57	72.39
Nnge	73.00	67.71	69.36
RandomSubSpace	69.60	68.45	68.81
RandomCommittee	70.60	67.86	68.71
LMT	71.60	65.89	67.67
Bagging	69.60	66.65	67.57
SimpleLogistic	71.60	65.32	67.29
RBFNetwork	75.60	62.19	66.38
LogitBoster	69.40	64.72	66.18
RealAdaBoost	68.20	65.12	66.08
AttributeSelectedClassifier	67.40	64.04	65.09
J48	66.80	63.43	64.49
AdaBoostM1	69.80	61.11	63.82
InterativeClassifierOptimizer	65.20	62.75	63.52
PART	65.60	62.50	63.47
SimpleCart	63.40	62.96	63.10
BayesNet	66.80	61.10	62.88
DecisionTable	62.40	61.82	62.00
FilteredClassifier	63.60	60.80	61.67
IB1	61.60	61.28	61.38
Kstar	61.40	61.31	61.34
Jrip	66.60	58.94	61.33
IBK	61.80	61.04	61.28
ClassificationViaRegression	64.20	59.10	60.69
REPTree	60.80	59.96	60.22
Ridor	62.60	57.04	58.78
LWL	61.80	57.20	58.64
RandomTree	62.20	56.99	58.62
OneR	62.80	56.27	58.31
DecisionStump	59.80	55.87	57.10
ConjunctiveRule	58.00	53.95	55.22
RandomizableFilteredClassifier	55.40	54.00	54.44
NaiveBayesMultinomialText	50.00	50.00	50.00
ZeroR	50.00	50.00	50.00

Qëllimi i eksperimenteve ka qenë të identifikohet roli i komponentit linguistik dhe i karakteristikave të ndryshme të parapërpunimit në Weka. Në këto eksperimente mbi secilin korpus është aplikuar vetëm faza e parë e komponentit linguistik, pra të gjithë fjalët janë vendosur në shkronja të vogla dhe janë fshirë të gjitha lidhëzat, karakteret speciale, shenjat e pikësimit, dhe numrat. Faza e dytë nuk është aplikuar, pra nuk kemi konvertuar çdo fjalë të opinionit në rrënjën përkatëse. Më tej, në krijimin e file .arff në Weka është aplikuar një nga 6 konfigurimet e përcaktuara në çështjen 6.1.5.

Në Tabelën 6.10 deri në Tabelën 6.16 paraqiten rezultatet e eksperimenteve të çdo algoritëm për çdo korpus në term të përqindjes së instancave të klasifikuara në mënyrë korrekte. Në kolonën e parë g_1 ose g_2 janë dhënë rezultatet e algoritmit për atë korpus për dy grupet e para të eksperimenteve. Në kolonat a deri në f janë paraqitur rezultatet e këtij grupi të tretë eksperimentesh për secilën nga për 6 konfigurimet e përcaktuar në çështjen 6.1.5. Me të zezë është theksuar rezultati më i mirë për atë korpus.

Tabela 6.10 Rezultatet për korpusin C1

	C_1						
Karakteristika	**g_1**	**a**	**b**	**c**	**d**	**e**	**f**
NaiveBayes Multinomial	87.00	92.00	92.00	93.00	93.00	93.00	93.00
SGD	88.00	92.00	92.00	92.00	92.00	93.00	93.00
RBFNetwork	49.00	93.00	93.00	51.00	51.00	95.00	51.00
Bayesian LogisticRegretion	89.00	91.00	90.00	89.00	89.00	89.00	89.00
Logistic	**94.00**	92.00	92.00	93.00	93.00	93.00	93.00
HyperPipes	92.00	92.00	92.00	91.00	91.00	92.00	92.00
SMO	88.00	90.00	90.00	92.00	92.00	91.00	91.00
Random Forest	86.00	83.00	81.00	88.00	90.00	90.00	88.00
VotedPerceptron	86.00	82.00	82.00	81.00	82.00	80.00	82.00
SimpleLogistic	84.00	80.00	80.00	82.00	81.00	80.00	82.00
J48	87.00	72.00	72.00	70.00	70.00	70.00	70.00
IBK	60.00	63.00	63.00	63.00	63.00	67.00	67.00

Tabela 6.11 Rezultatet për korpusin C2

	C_2						
Karakteristika	**g_1**	**a**	**b**	**c**	**d**	**e**	**f**
NaiveBayes Multinomial	89.00	95.00	95.00	94.00	97.00	94.00	95.00
SGD	84.00	89.00	89.00	89.00	89.00	87.00	87.00
RBFNetwork	88.00	50.00	92.00	97.00	97.00	97.00	97.00
Bayesian LogisticRegretion	85.00	92.00	91.00	90.00	94.00	91.00	93.00
Logistic	84.00	94.00	94.00	84.00	84.00	87.00	87.00
HyperPipes	92.00	93.00	93.00	**98.00**	**98.00**	**98.00**	**98.00**
SMO	82.00	87.00	87.00	82.00	82.00	81.00	81.00
Random Forest	76.00	80.00	83.00	67.00	78.00	79.00	77.00
VotedPerceptron	76.00	88.00	84.00	87.00	88.00	88.00	83.00
SimpleLogistic	74.00	68.00	65.00	70.00	68.00	67.00	65.00
J48	67.00	57.00	57.00	57.00	57.00	59.00	59.00
IBK	65.00	64.00	64.00	65.00	65.00	65.00	65.00

Tabela 6.12 Rezultatet për korpusin C3

	C_3						
Karakteristika	**g_1**	**a**	**b**	**c**	**d**	**e**	**f**
NaiveBayes Multinomial	77.00	82.00	82.00	84.00	87.00	86.00	89.00
SGD	75.00	83.00	83.00	82.00	82.00	83.00	83.00
RBFNetwork	66.00	90.00	90.00	**92.00**	**92.00**	86.00	86.00
Bayesian LogisticRegretion	77.00	81.00	82.00	86.00	85.00	87.00	85.00
Logistic	66.00	82.00	82.00	90.00	90.00	82.00	82.00
HyperPipes	67.00	80.00	80.00	88.00	88.00	82.00	82.00
SMO	74.00	82.00	82.00	81.00	81.00	84.00	84.00
Random Forest	*76.00*	79.00	77.00	78.00	79.00	79.00	82.00
VotedPerceptron	74.00	79.00	77.00	80.00	78.00	82.00	80.00
SimpleLogistic	66.00	69.00	68.00	71.00	68.00	68.00	68.00
J48	58.00	76.00	76.00	65.00	65.00	65.00	65.00
IBK	62.00	60.00	60.00	56.00	56.00	62.00	62.00

Tabela 6.13 Rezultatet për korpusin C4

	C_4						
Karakteristika	**g_1**	**a**	**b**	**c**	**d**	**e**	**f**
NaiveBayes Multinomial	79.00	83.00	83.00	87.00	89.00	87.00	87.00
SGD	75.00	76.00	76.00	76.00	76.00	71.00	71.00
RBFNetwork	86.00	88.00	88.00	90.00	90.00	89.00	89.00
Bayesian LogisticRegretion	74.00	72.00	72.00	70.00	72.00	71.00	71.00
Logistic	82.00	86.00	86.00	88.00	88.00	85.00	85.00
HyperPipes	85.00	88.00	88.00	**92.00**	**92.00**	**92.00**	**92.00**
SMO	75.00	71.00	71.00	73.00	73.00	74.00	74.00
Random Forest	65.00	74.00	67.00	70.00	68.00	67.00	70.00
VotedPerceptron	71.00	70.00	76.00	68.00	74.00	72.00	75.00
SimpleLogistic	63.00	67.00	67.00	69.00	71.00	69.00	72.00
J48	60.00	53.00	53.00	58.00	58.00	58.00	58.00
IBK	60.00	62.00	62.00	68.00	68.00	65.00	65.00

Tabela 6.14 Rezultatet për korpusin C5

	C_5						
Karakteristika	**g_1**	**a**	**b**	**c**	**d**	**e**	**f**
NaiveBayes Multinomial	85.00	86.00	84.00	87.00	90.00	89.00	89.00
SGD	81.00	80.00	80.00	87.00	87.00	84.00	84.00
RBFNetwork	89.00	51.00	93.00	**93.00**	**93.00**	**93.00**	**93.00**
Bayesian LogisticRegretion	78.00	76.00	80.00	78.00	82.00	77.00	81.00
Logistic	79.00	80.00	80.00	86.00	86.00	89.00	89.00
HyperPipes	80.00	81.00	81.00	90.00	90.00	90.00	90.00
SMO	78.00	76.00	76.00	79.00	79.00	81.00	81.00
Random Forest	74.00	74.00	78.00	77.00	76.00	82.00	75.00
VotedPerceptron	75.00	76.00	73.00	78.00	82.00	77.00	79.00
SimpleLogistic	71.00	61.00	61.00	63.00	63.00	66.00	66.00
J48	62.00	67.00	67.00	64.00	64.00	75.00	75.00
IBK	62.00	50.00	50.00	58.00	58.00	59.00	59.00

Tabela 6.15 Rezultatet për koprusin C6

	C_6						
Karakteristika	**g_1**	**a**	**b**	**c**	**d**	**e**	**f**
NaiveBayes Multinomial	78.6	76.80	78.40	77.60	79.40	78.00	**78.80**
SGD	77.4	75.00	75.00	75.20	75.20	75.60	75.60
RBFNetwork	73.4	76.20	76.20	76.20	76.20	74.40	74.40
Bayesian LogisticRegretion	77.0	75.60	76.00	75.60	77.20	74.80	75.80
Logistic	66.6	68.40	68.40	67.40	67.40	64.40	64.40
HyperPipes	55.8	58.60	58.60	62.20	62.20	60.00	60.00
SMO	76.0	71.60	71.60	71.60	71.60	70.80	70.80
Random Forest	74.2	77.00	59.60	77.40	76.80	77.40	77.80
VotedPerceptron	69.2	70.20	70.00	72.20	71.60	70.80	73.00
SimpleLogistic	66.4	68.80	68.80	68.20	68.20	67.00	67.00
J48	61.2	63.80	63.80	65.80	65.80	65.80	65.80
IBK	60.4	59.20	59.20	60.40	60.40	60.00	60.00

Tabela 6.16 Rezultatet për korpusin C17

	C_17						
Karakteristika	**g_2**	**a**	**b**	**c**	**d**	**e**	**f**
NaiveBayes Multinomial	74.11	74.22	75.78	72.89	73.89	72.89	73.78
SGD	72.44	72.89	72.89	74.56	74.56	74.22	74.22
RBFNetwork	70.44	71.89	71.89	69.22	69.22	69.78	69.78
Bayesian LogisticRegretion	73.89	73.11	73.89	72.22	73.11	72.67	73.11
Logistic	59.44	61.33	61.33	59.56	59.56	60.44	60.44
HyperPipes	50.67	49.67	49.67	49.67	49.67	49.67	49.67
SMO	*71.67*	72.33	72.33	72.00	72.00	71.78	71.78
Random Forest	75.56	**77.11**	75.44	74.67	75.00	75.11	75.56
VotedPerceptron	70.89	71.56	73.33	70.56	70.00	70.22	68.78
SimpleLogistic	68.22	69.78	69.78	68.44	68.44	69.11	69.11
J48	60.00	62.11	62.11	61.33	61.33	61.33	61.33
IBK	60.67	58.11	58.11	57.78	57.78	58.11	58.11

Për korpusin C_1, algoritmi me performancë më të lartë mbetet algoritmi Logistic, por performanca e tij ulet me aplikimin e karakteristikave të reja të papapërpunimit.

Për korpusin C_2, algoritmi me performancë më të lartë mbetet algoritmi Hyper Pipes, por performanca e tij rritet me aplikimin e karakteristikave të reja të papapërpunimit.

Për korpusin C_3, algoritmi me performancë më të lartë është algoritmi RBF Network, por performanca e tij rritet me aplikimin e karakteristikave të reja të papapërpunimit.

Për korpusin C_4, algoritmi me performancë më të lartë është algoritmi Hyper Pipes, por performanca e tij rritet me aplikimin e karakteristikave të reja të papapërpunimit.

Për korpusin C_5, algoritmi me performancë më të lartë mbetet algoritmi RBF Network, por performancae tij rritet me aplikimin e karakteristikave të reja të papapërpunimit.

Për korpusin C_6, algoritmi me performancë më të lartë mbetet algoritmi Naive Bayes Multinomial, por performanca e tij rritet me aplikimin e karakteristikave të reja të papapërpunimit.

Për korpusin C_17, algoritmi me performancë më të lartë është algoritmi Random Forest, por performanca e tij rritet me aplikimin e karakteristikave të reja të papapërpunimit.

Në Tabelën 6.17 për çdo algoritëm është llogaritur një vlerë mesatare e përqindjes së instancave të klasifikuara në mënyrë korrekte për çdo rast parapërpunimi të përdorur dhe vlera mesatare totale. Me të zezë është theksuar rezultati i algoritmit më të mirë dhe me shkronja të pjerrëta rezultati më i mirë për çdo konfigurim karakteristikash. Nëse bazohemi në vlerën mesatare totale, algoritmi Naive Bayes Multinomial ka performancën më të lartë.

Tabela 6.17 Mesatarja e rezultateve për karakteristikë dhe mesatarja totale

				Mesatare				***Mesatare totale***
Karakteristika	**g_1_2**	**a**	**b**	**c**	**d**	**e**	**f**	
NaiveBayes Multinomial	*81.39*	*84.15*	84.31	*85.07*	***87.04***	85.7	*86.51*	**84.88**
SGD	78.98	81.13	81.13	82.25	82.25	81.12	81.12	81.14
RBFNetwork	74.55	74.30	*86.30*	81.20	81.20	*86.31*	80.03	80.56
Bayesian Logistic Regretion	79.13	80.10	80.70	80.12	81.76	80.35	81.13	80.47
Logistic	75.86	80.53	80.53	81.14	81.14	80.12	80.12	79.92
HyperPipes	74.64	77.47	77.47	81.55	81.55	80.52	80.52	79.10
SMO	77.81	78.56	78.56	78.66	78.66	79.08	79.08	78.63
Random Forest	75.25	77.73	74.43	76.01	77.54	78.50	77.91	76.77
VotedPerceptron	74.58	76.68	76.48	76.68	77.94	77.15	77.25	76.68
SimpleLogistic	70.37	69.08	68.51	70.23	69.66	69.44	69.87	69.60
J48	65.03	64.42	64.42	63.02	63.02	64.88	64.88	64.24
IBK	61.44	59.47	59.47	61.17	61.17	62.30	62.30	61.05

Nga rezultatet mund të themi që algoritmet kanë performancë më të mirë në rastin kur korpusi nuk ka kaluar në fazën e dytë të komponentit linguistik, që e kthen çdo fjalë në rrënjën e saj. Në terma mesatare të rezultateve, gjashtë nga algoritmet kanë performancë më të mirë në rastin kur përdoret TF-IDF dhe n-gram me vlerat min=1 dhe max=2, katër nga algoritmet kanë performancë më të mirë në rastin kur përdoret n-gram me vlerat min=1 dhe max=3 dhe dy nga algoritmet kur përdoret WordTokenizer.

Edhe në këtë rast kemi realizuar eksperimentin e vlerësimit të kryqëzuar për të identifikuar ndonjë ndryshim statistikor në performancën e këtyre algoritmeve. Si algoritëm bazë është zgjedhur algoritmi Naïve Bayes Multinomial. Eksperimenti është realizuar duke përdorur të gjithë korpuset, 10 përsëritje, dhe 10 *cross-validation*. Rezultatet e eksperimenti janë paraqitur në Tabelën 6.18.

Në rezultatet e pothuajse të të gjitha algoritmeve kemi praninë e "*", që tregon që këto algoritme kanë performancë më të ulët se algoritmi bazë Naïve Bayes Multinomial. Gjithashtu, nuk kemi praninë e një "v", që do të thotë asnjë nga algoritmet nuk ka performuar më mirë se algoritmi bazë Naïve Bayes Multinomial. Në rreshtin e fundit të tabelës, për çdo algoritëm kemi një vlerë të formës (a/b/c), që përcakton: a- numrin e herëve që algoritmi ka performuar më mirë, b-numrin e herëve që algoritmi ka performuar njëlloj dhe c-numrin e herëve që algoritmi ka performuar më keq se algoritmi bazë Naïve Bayes Multinomial. Tri algoritmet Simple Logistic, J48 dhe IBk në të gjitha korpuset kanë performuar më keq se algoritmi bazë Naïve Bayes Multinomial. Dy algoritmet RBF Network dhe Bayesian Logistic Regression në gjashtë korpuse nuk kanë diferencë në performancë kundrejt algoritmit bazë Naïve Bayes Multinomial, kurse në një korpus performojnë më keq se ai. Midis tri algoritmeve RBF Network, Bayesian Logistic Regression dhe Naïve Bayes Multinomial nuk mund të themi që ka ndonjë diference statistikore në performancë.

Në Tabelën 6.19 paraqitet një përmbledhje e rezultateve të eksperimentit të mësipërm. Numri jashtë kllapave rrumbullake përcakton numrin e herëve që algoritmi i përcaktuar në kolonë ka performancë më të mirë se algoritmi i përcaktuar në rresht. Vlera zero përcakton që algoritmi i përcaktuar në kolonë nuk ka performuar asnjë herë më mirë kundrejt atij të përcaktuar në rresht. Numri brenda kllapave rrumbullake tregon numrin e herëve që algoritmi i përcaktuar në kolonë ka fituar kundrejt atij të përcaktuar në rresht.

Më tej, në Tabelën 6.20 tregohet rezultati i testit të renditjes të rezultateve. Në këtë test tregohet numri i herëve totale që një algoritëm ka performuar më mirë ose më keq se të gjithë algoritmet e tjera. Në kolonën > përcaktohet numri i herëve që algoritmi ka performuar më mirë, në kolonën < përcaktohet numri i herëve që algoritmi ka performuar më keq dhe në kolonën >-< përcaktohet diferenca e dy rezultateve të para. Algoritmi Naïve Bayes Multinomial është algoritmi i renditur në vend të parë, i ndjekur nga RBF Network në vend të dytë. Algoritmet e tjera kanë vlerë të ulët të diferencës, madje dhe negative për pesë nga algoritmet. Kjo përcakton që këto algoritme kanë performancë të ulët.

Tabela 6.18 Rezultatet e eksperimentit për vlerësimin e kryqëzuar

Korpus	***Naïve Bayes Multinomial***	***SGD***	***Bayesian Logistic Regression***	***Hyper Pipes***	***RBF Network***	***Logistic***	***SMO***	***Voted Perceptron***	***Random Forest***	***Simple Logistic***	***J48***	***IBk***
C_1	92.10 |	91.4	87.6	91.5	95.1	92.5	89.7	83.60*	85.3	77.60*	74.30*	63.20*
C_2	96.20 |	88.30*	91.7	97.9	97.4	90.5	83.60*	86.70*	80.70*	66.60*	58.80*	63.40*
C_3	85.30 |	83.1	86.9	87.4	90.9	84.5	80.5	78.3	79.4	68.90*	67.50*	57.40*
C_4	88.80 |	76.50*	76.70*	93	93.3	88.2	75.00*	72.90*	71.50*	65.30*	58.90*	66.60*
C_5	87.80 |	81.1	82.5	90	93.1	86.2	80.2	76.40*	78.8	66.10*	65.40*	57.80*
C_6	79.06 |	74.76*	76.36	61.72*	76.34	67.84*	71.44*	73.02*	77.68	67.40*	64.28*	60.12*
C_7	73.72 |	73.44	72.33	49.87*	68.86*	59.11*	71.28	71.44	75.19	67.60*	61.21*	58.72*
	(v/ /*) |	(0/4/3)	(0/6/1)	(0/5/2)	(0/6/1)	(0/5/2)	(0/4/3)	(0/2/5)	(0/5/2)	(0/0/7)	(0/0/7)	(0/0/7)

Tabela 6.19 Përmbledhje e rezultateve të eksperimentit të kryqëzuar

a	***b***	***c***	***d***	***e***	***f***	***g***	***h***	***i***	***j***	***k***	***L***	***|(No. of datasets where [col] >> [row])***
-	0 (0)	1 (0)	4 (0)	5 (0)	1 (0)	0 (0)	0 (0)	1 (0)	0 (0)	0 (0)	0 (0)	|a = (1) NaiveBayesMultinomial
7 (3)	-	5 (0)	5 (3)	6 (3)	5 (1)	0 (0)	0 (0)	2 (0)	0 (0)	0 (0)	0 (0)	|b = (2) SGD
6 (1)	2 (0)	-	5 (1)	5 (3)	3 (1)	1 (0)	0 (0)	2 (1)	0 (0)	0 (0)	0 (0)	|c = (3) BayesianLogisticRegression
3 (2)	2 (2)	2 (2)	-	6 (2)	3 (2)	2 (2)	2 (2)	2 (2)	2 (2)	2 (1)	1 (1)	|d = (4) HyperPipes
2 (1)	1 (1)	2 (0)	1 (0)	-	0 (0)	1 (0)	1 (0)	2 (1)	0 (0)	0 (0)	0 (0)	|e = (5) RBFNetwork
6 (2)	2 (2)	4 (2)	4 (0)	7 (2)	-	2 (1)	2 (1)	2 (2)	1 (1)	1 (0)	0 (0)	|f = (6) Logistic
7 (3)	7 (2)	6 (2)	5 (3)	6 (3)	5 (1)	-	3 (0)	2 (2)	0 (0)	0 (0)	0 (0)	|g = (7) SMO
7 (5)	7 (1)	7 (0)	5 (4)	6 (5)	5 (2)	4 (0)	-	5 (0)	0 (0)	0 (0)	0 (0)	|h = (8) VotedPerceptron
6 (2)	5 (1)	5 (1)	5 (3)	5 (5)	5 (2)	5 (0)	2 (0)	-	0 (0)	0 (0)	0 (0)	|i = (9) RandomForest
7 (7)	7 (7)	7 (6)	5 (5)	7 (6)	6 (5)	7 (4)	7 (2)	7 (5)	-	0 (0)	1 (0)	|j = (10) SimpleLogistic
7 (7)	7 (7)	7 (7)	5 (5)	7 (7)	6 (5)	7 (7)	7 (4)	7 (4)	7 (1)	-	2 (0)	|k = (11) J48
7 (7)	7 (6)	7 (7)	6 (5)	7 (7)	7 (6)	7 (6)	7 (6)	7 (6)	6 (4)	5 (0)	-	|l = (12) IBk

Tabela 6.20 Rezultati i testit të renditjes të rezultateve të eksperimentit

Rezultatet	>-<	>	<
NaiveBayesMultinomial	40	40	0
RBFNetwork	38	43	5
BayesianLogisticRegression	18	27	9
SGD	13	29	16
Logistic	8	25	17
RandomForest	5	23	18
HyperPipes	5	29	24
SMO	-2	20	22
VotedPerceptron	-12	15	27
SimpleLogistic	-53	8	61
J48	-67	1	68
IBk	-73	1	74

6.1.7. Vlerësimi eksperimental i rrjetit neural

Në këtë pjesë prezantohen modelet e rrjeve neurale të implementuara për të klasifikuar opinionet në gjuhën shqipe, mbi bazën e polaritetit të ndjenjës pozitive apo negative që ato shprehin. Modelet e rrjeteve neurale kërkojnë sasi të mëdha të dhënash për trajnim, prandaj dhe kemi zgjedhur të përdorim vetëm korpusin C_17 në këtë rast. Korpusi është parapërpunuar vetëm duke përdorur fazën e parë të komponentit linguistik dhe është ndarë në dy pjesë: në korpusin i trajnimit dhe në korpusin i testimit. Për të implementuar rrjetin neural kemi përdorur Keras (Chollet, 2015) dhe Tensorflow (Abad, et al., 2015). Janë implementuar dy modele: një model i rrjetit neural *bag-of-words* dhe një model rrjeti neural Convolutional Neural Network (CNN). Në vazhdim do të diskutohet arkitektura e rrjetit të përdorur dhe vlerësimi eksperimental i tij.

6.1.7.1. Modeli bag-of-words

Modeli i parë i implementuar është një model që e trajton korpusin e opinioneve tekst si një set fjalësh. Para se korpusi të përdoret në trajnimin dhe testimin e modelit duhet të konvertohen në vektorë karakteristikash që ka aq dimensione sa karakteristika unike kemi në korpus. Për të realizuar këtë gjë janë përdorur funksione dhe klasa të përcaktuara në Scikit (Pedregosa, et al., 2011). Fillimisht është krijuar një matricë për të dhënat dhe një matricë për etiketat. Dimensionet e matricës së të dhënave janë 900 aq sa është dhe numri i opinioneve në korpusin e përdorur, dhe 27652 aq sa është numri maksimal i karakteristikave në korpus. Një element i matricës ka vlerën 0 ose 1 për çdo karakteristikë në varësi nëse ajo ndodhet apo jo në tekstin e një opinioni të caktuar. Më tej, të dhënat janë vektorizuar duke përdorur *CountVectorizer* me 1-gram. Matrica e etiketave përmban një

rresht për çdo opinion tekst në korpus me dy kolona, në njërën nga kolonat ruhet vlera 1 që përcakton etiketën me të cilën është etiketuar opinioni dhe në kolonën tjetër ruhet vlera 0. Matrica e etiketave është enkoduar me *one-hot encoding*.

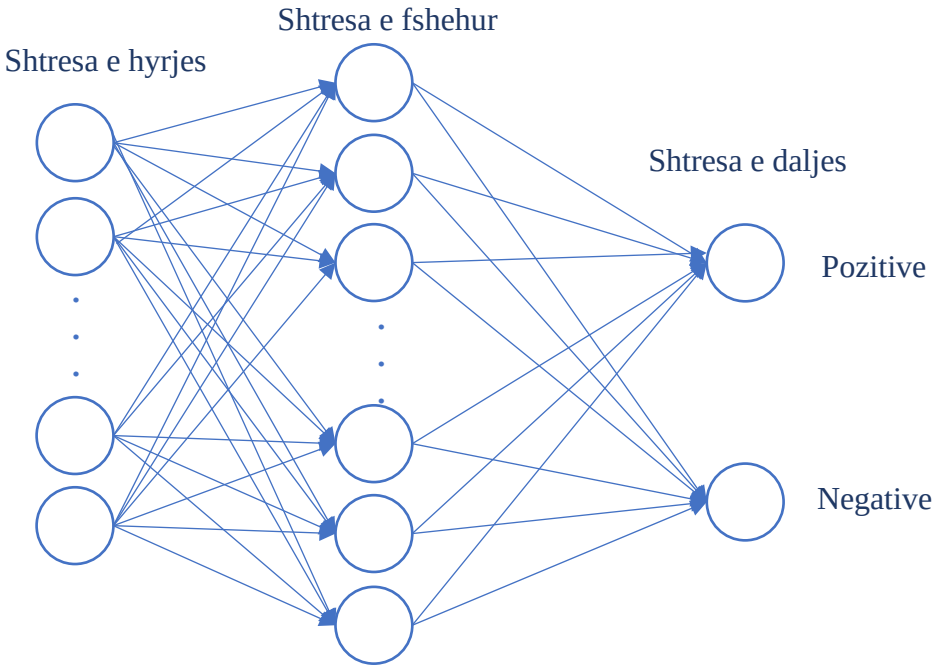


Figura 6.5 Arkitektura e rrjetit bag-of-words

Në Figurën 6.5 paraqitet arkitektura e rrjetit të përdorur dhe në Tabelën 6.21 parametrat e përdorura. Arkitektura e rrjetit është një arkitekturë me tri shtresa. Shtresa e hyrjes ka aq nyje sa numri i karakteristikave të korpusit, shtresa e fshehur ka 900 nyje dhe shtresa e daljes ka 2 nyje që përfaqësojnë etiketën pozitive dhe negative me të cilën etiketohet opinioni. Për të stabilizuar dhe për të patur rezultate më të mira është përdorur dropout = 0.5 dhe modeli është ekzekutuar për 10 epoka. Pas 10 epokash performanca e modelit nuk përmirësohet më dhe modeli kalon në mbingopje.

Tabela 6.21 Specifikimet e modelit bag-of-words

	Specifikimet		**Specifikimet**
Shtresa e hyrjes	27652 nyje	**Optimizer**	
Dropout	0.5	Adam	lr=0.0001
Shtresa e fshehur	Dense	Loss	Categorical cross entropy
Nyje	900	Metrics	Accuracy
Activation	ReLu	**Batch size**	100
Shtresa e daljes	Dense	**verbose**	1
Nyje	1 pozitive & 1 negative	**Epoka**	10
Activation	Softmax	**Ndarja për vlerësim**	0.2

Në Tabelën 6.22 paraqiten rezultatet e vlerësimit të modelit. Modeli ka arritur vlerë të F1-score prej 82% në terma mesatar. Rezultatet tregojnë që modeli ka performancë të mirë.

Tabela 6.22 Rezultate eksperimentale të modelit bag-of-words

	Precision	***Recall***	***F1-score***
Negative	0.81	0.81	0.81
Positive	0.84	0.84	0.84
Mesatare totale	0.82	0.82	0.82

6.1.7.2. Modeli CNN

Në implementimin e këtij modeli kemi përdorur një rrjet neural CNN. Në Figurën 6.6 paraqitet arkitektura e modelit dhe në Tabelën 6. 23 paraqiten parametrat e përdorura. Në model është përdorur rrjeti CNN1, që është një rrjet shumë eficient për të nxjerrë karakteristika nga segmente të dhënash me një madhësi fikse. Dallimi midis rrjetit CNN1 dhe CNN2 është mënyra se si deduktuesi i karakteristika i identifikon ato në të dhëna.

Si shtresë hyrje është përdorur një shtresë *embeded* duke përdorur vektorët e paratrajnuar nga Fasttext (2018). Në këtë model të dhënat nuk mund të jenë një set të dhënash por një sekuencë të dhënash. Hapi i parë është përgatitja e të dhënave që do të përdoren për trajnimin dhe testimin e modelit. Fjalori i të dhënave është zgjeruar duke përdorur fjalë nga modeli *embedding* dhe më pas është ndërtuar fjalori i vektorizuar ku çdo fjalë përfaqësohet nga një numër. Pas kësaj krijohet matrica *embedding*.

Shtresa pasardhëse është një shtresë Conv1D me 100 neurone dhe ReLu. Në këtë shtresë përdoret një rregullator l2 dhe SpatialDropout1D, të cilat e detyrojnë modelin të mësojë individualisht karakteristikat dhe të rrisi performancën. Më pas aplikohet një shtresë GlobalMaxPooling1D për të përzgjedhur karakteristikat më të rëndësishme. Shtresa e fundit është një shtresë Dense me softmax activation që gjeneron në dalje probabilitetin e dy etiketave, pozitive dhe negative.

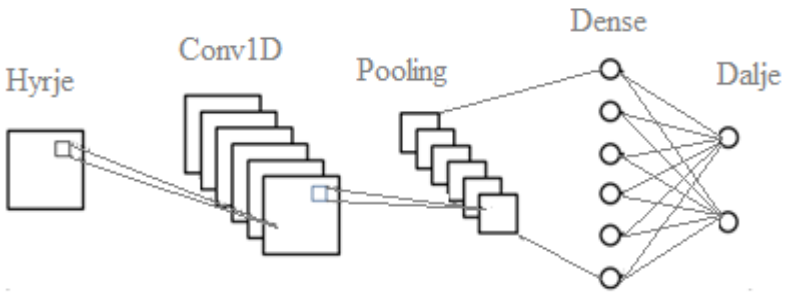


Figura 6.6 Arkitektura e rrjetit CNN

Tabela 6.23 Specifikimet e modelit CNN

	Specifikimet		**Specifikimet**
Shtresa e hyrjes	Embeddings	**Optimizer**	
Shtresa e fshehur	Conv1D	Adam	lr=0.0001
Nyje	900	Loss	Categorical cross entropy
Activation	ReLu	Metrics	Accuracy
Activity regularizer	l2 (0.0002) e SpatialDropout1D	**Batch size**	100
Shtesa Pooling	GlobalMaxPooling1D	**Verbose**	1
Shtresa e daljes	Dense	**Epoka**	10
Nyje	2 (1 pozitive & 1 negative)	**Ndarja për vlerësim**	0.2
Activation	Softmax		

Edhe në këtë rast modeli është trajnuar për 10 epoka dhe pas 10 epokash performanca e modelit nuk përmirësohet më, por modeli kalon në mbingopje.

Saktësia e këtij modeli është 73%. Rezultatet tregojnë se ky model performon më keq se modeli i parë.

6.1.8. Përfundime

Në këtë pjesë është realizuar një vlerësim eksperimental i algoritmeve të të mësuarit e automatizuar të kontrolluar të implementuar në platformën Weka dhe të dy modeleve të rrjeteve neurale të implementuar në platformat Keras dhe Tensorflor, për të realizuar detyrën e klasifikimit të opinioneve tekst në gjuhën shqipe në një nga dy klasat pozitive dhe negative. Në eksperimentet e realizuara është vlerësuar performanca e 50 algoritmeve MA dhe e dy rrjeteve neurale, modelit *bag-of-words* dhe modeli CNN duke përdorur korpuse nga tema të ndryshme dhe me numër të ndryshëm dokumentesh. Nga përfundimet e eksperimenteve në term mesatar mund të themi që algoritmet Naive Bayes Multinomial dhe RBF Network kanë performancën më të mirë nga 50 algoritmet e marra në konsideratë. Performanca e algoritmeve përmirësohet në përdorimin e TF-IDF dhe n-gram me vlerë min =1 dhe max=2. Modeli *bag-of-words* ka performancën më të mirë në përgjithësi. Vlerësimet eksperimentale të realizuara dhe të paraqitura në këtë pjesë janë publikuar në dy konferenca (Kote, Biba, & Trandafili, 2018; Kote & Biba, 2018), dhe një revistë (Kote & Biba, 2020).

6.2. Etiketimi i pjesëve të ligjëratës dhe temëzimi në gjuhën shqipe

Etiketuesi morfologjik dhe temëzuesi që trajtohet në këtë pjesë bazohet në trajnimin e një modeli të parserit Turku Pipeline të trajtuar në çështjen 5.7. Parseri i përdorur është një parser i cili bazohet në të mësuarin e automatizuar të kontrolluar. Për të realizuar këtë

etiketues është krijuar një korpus i etiketuar duke përshtatur skemën e etiketimit të Universal Dependencies (UD) për gjuhën shqipe.

6.2.1. Përzgjedhja e korpusit

Sistemet që përdorin të mësuarin e automatizuar të kontrolluar për të etiketuar pjesët e ligjëratës ndikohen në performancën e tyre nga korpusi që përdoret për të mësuar sistemi. Në dijeninë tonë, në gjuhën shqipe nuk ekziston një korpus i etiketuar me etiketat për pjesët e ligjëratës me akses publik për t'u përdorur. Për këtë arsye kemi krijuar një korpus të etiketuar me etiketën për pjesët e ligjëratës, etiketat për karakteristikat morfologjike dhe temën e fjalës. Korpusi përmban 117,688 fjalë (6,640 fjali) të mbledhura nga burime të ndryshme: fjali të zgjedhura nga dy romane, fjali të zgjedhura nga libra gramatikorë për të përfshirë fjali komplekse dhe me karakteristika morfologjike të zgjeruara, fjali të përzgjedhura nga korpusi i gjuhës shqipe të Leipzig Corpora Collection (Goldhahn, et al., 2012) dhe fjali të thjeshta të krijuara duke përdorur 66,911 fjalë (17,042 fjali) nga një listë fjalësh të etiketuara nga projekti UniMorph (Kirov, et al., 2018).

Në Figurën 6.7 paraqiten hapat që janë ndjekur për përgatitjen e korpusit final, për trajnimin dhe vlerësimin e modelit të sistemit për etiketimin e pjesëve të ligjëratës, karakteristikave morfologjike dhe temëzim.

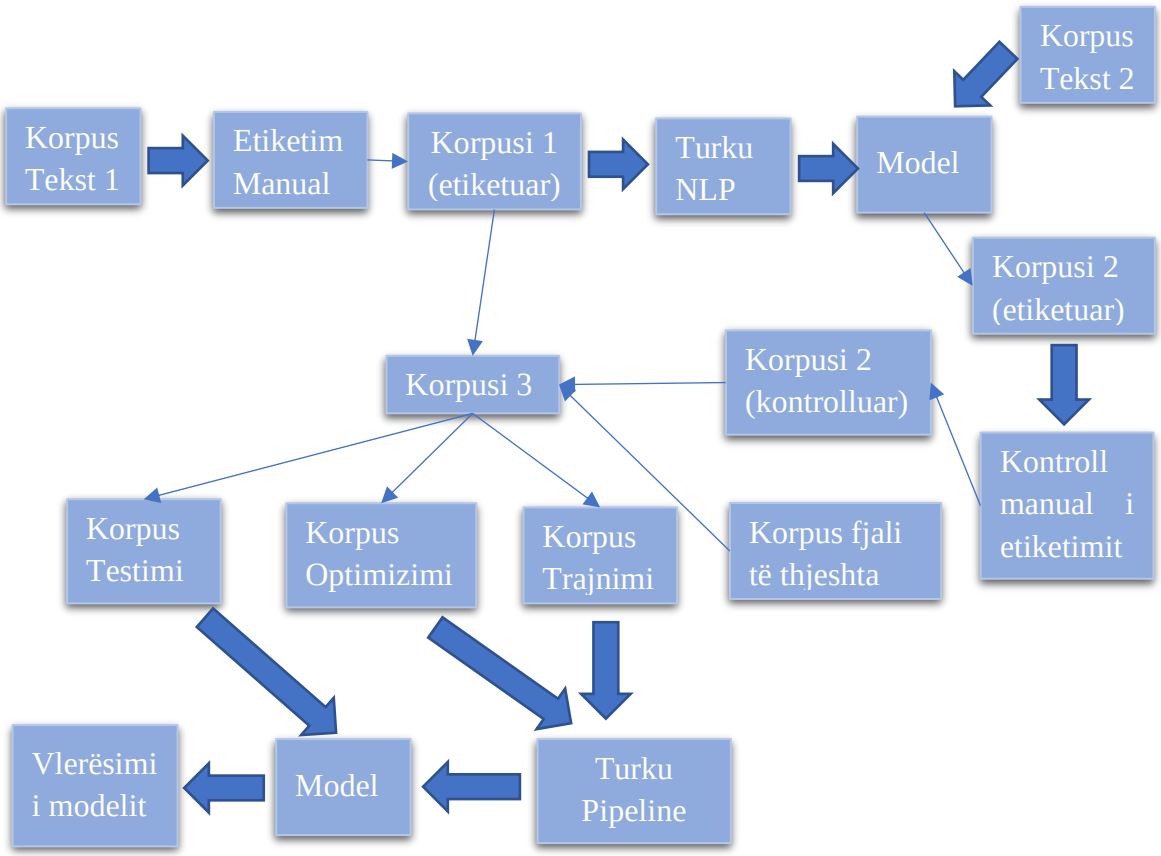


Figura 6.7 Skema për krijimin e korpusit final dhe trajnimin dhe vlerësimin e sistemit

Si pikënisje për krijimin e korpusit është shfrytëzuar korpusi i etiketuar nga Salavaçi dhe Biba (2012), i cili është konvertuar në skemën e përdorur nga UD duke përdorur një skript automatik. Për çdo etiketë të përcaktuar nga Salavaçi dhe Biba (2012) është përcaktuar etiketa e pjesëve të ligjëratës dhe etiketat e karakteristikave morfologjike sipas skemës UD. Më pas, korpusi është rishikuar për gabime të mundshme në etiketim. Duke qenë se korpusi nuk përmban temat e fjalëve ato janë shtuar manualisht. Ky korpus përmban 7,374 fjalë (476 fjali) të përzgjedhura nga dy libra artistik: Kadare (2011) dhe Fojhtvanger (1999) dhe nga libri gramatikor i Lafe et al. (1979).

Ky korpus është përdorur për të trajnuar një etiketues duke përdorur parserin Turku Pipeline, që është përdorur më tej për të etiketuar pjesën tjetër të korpusit të përzgjedhur nga Leipzig Corpora Collection (Goldhahn, et al., 2012). Pas etiketimin automatik ky korpus është kontrolluar manualisht nga dy persona, njëri nga të cilët specialist gjuhësor për të identifikuar dhe për të rregulluar gabime në etiketim. Më tej, korpusi është zgjeruar dhe me fjalitë e krijuara artificialisht nga lista e fjalëve të etiketuara nga projekti UniMorph (Kirov, et al., 2018). Ky korpus i zgjeruar është përdorur për të trajnuar një model të ri.

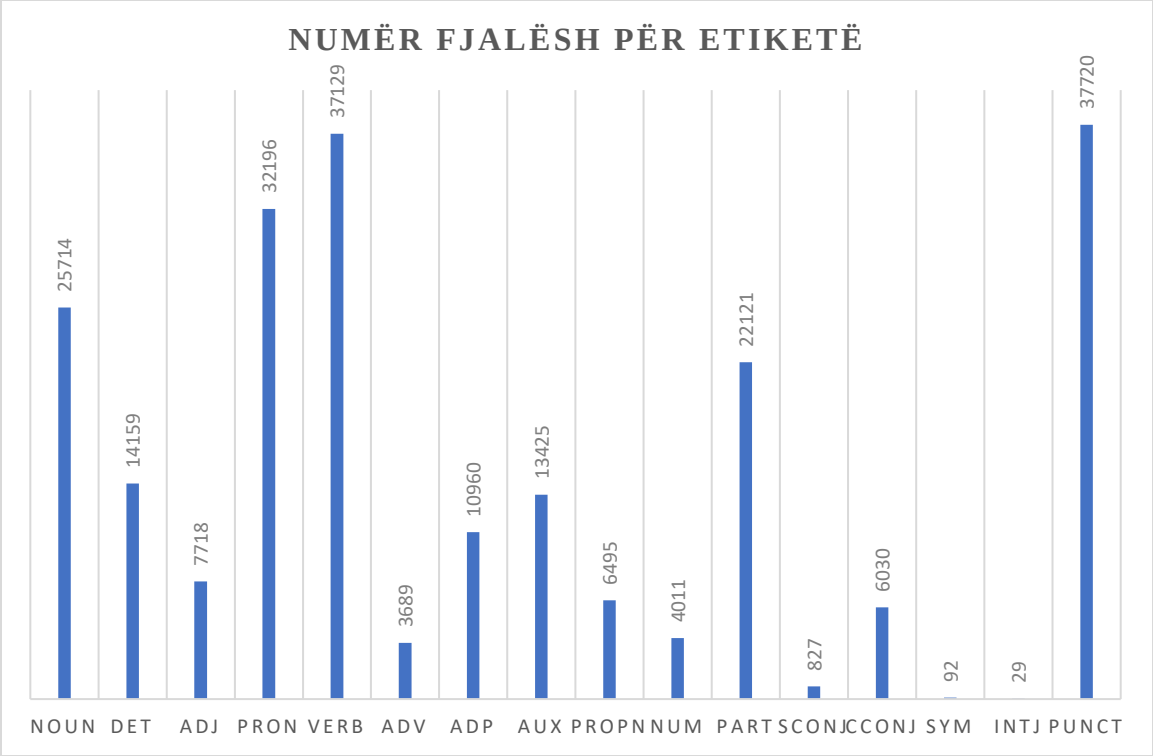
Korpusi i zgjeruar është ndarë në tri pjesë: korpusi i trajnimit, korpusi i zhvillimit dhe korpusi i testimit. Fjalitë nga i njëjti dokument janë përfshirë vetëm në njërin nga korpuset për të shmangur mbivendosje leksikore të paqëllimshme. Të gjithë fjalitë e krijuara artificialisht janë përfshirë në korpusin e trajnimit duke qenë se janë fjali të thjeshta, një përemër vetor si kryefjalë dhe folje, për të mos ndikuar në rritjen artificiale të performancës së etiketuesit në testim.

Në Tabelën 6.24 paraqiten statistika për ndarjen e korpusit.

Tabela 6.24 Statistika të ndarjes së korpusit

	Trajnim	***Zhvillim***	***Testim***
Fjalë			
Libra, web, Wikipedia	93,621	11,375	12,690
Fjali nga UniMorph	66,911	—	—
Totali	**160,532**	**11,375**	**12,690**
Fjali			
Libra, web, Wikipedia	5,324	612	708
Fjali nga UniMorph	17,042	—	—
Total	**22,366**	**612**	**708**

Në Grafikun 6.1 paraqiten statistika në lidhje me frekuencën e çdo etikete të pjesëve të ligjëratës në korpusin përfundimtar.



Grafiku 6.1 Statistika për shpërndarjen e etiketave të pjesëve së ligjëratës në korpus

6.2.2. Etiketimi i pjesëve të ligjëratës në gjuhën shqipe

Etiketimi i realizuar përfshin përcaktimin për çdo fjalë në korpus të etiketës që përcakton pjesën e ligjëratës dhe të etiketave që përcaktojnë karakteristikat morfologjike të formës së fjalës dhe përcaktimin e temës së fjalës. Korpusi i etiketuar është në përputhje me skemën e përcaktuar nga UD. Karakteristikë e kësaj skeme etiketimi është që etiketimi realizohet në nivel fjale dhe jo shprehje siç mund të nevojitet në disa raste në gjuhën shqipe. Më poshtë diskutohet me detaje mënyra e etiketimit. Korpusi është ruajtur në një file tekst të formatit CoNLL-U i enkoduar duke përdorur UTF-8. Formati i këtij file në mënyrë të detajuar është shpjeguar në çështjen 5.6.

Në Tabelën 6.25 janë specifikuar etiketat e pjesëve të ligjëratës të përdorura në korpus. Ne kemi përdorur 16 etiketa për pjesët e ligjëratës që përdoren në gramatikën e gjuhë shqipe, për simbolet dhe për shenjat e pikësimit. Janë përdorur etiketa të ndryshme për emrin e përgjithshëm dhe për emrin e përveçëm. Si emra të përveçëm janë etiketuar jo vetëm emrat e frymorëve por dhe emërtimet gjeografike, emërtimet e institucioneve, titujt e librave, etj. Gjithashtu edhe për foljet kemi përdorur dy etiketa, etiketa VERB për foljen kryesore dhe etiketa AUX nëse folja është përdorur në rolin e foljes ndihmëse.

Tabela 6.25 Lista e etiketave të përdorura për pjesët e ligjëratës

Pjesa e ligjëratës	***Etiketa***
Emri	NOUN
Emri i përveçëm	PROPN
Mbiemri	ADJ
Numërori	NUM
Përemri	PRON
Folja	VERB
Folja ndihmëse	AUX
Ndajfolja	ADV
Parafjala	ADP
Lidhëza	CCONJ / SCONJ
Nyja	DET
Pjesëza	PART
Pasthirrma	INTJ
Simbole	SYM
Shenja e pikësimit	PUNCT

Në Tabelat 6.26 janë specifikuar karakteristikat morfologjike të përdorura për etiketimin e emrit.

Tabela 6.26 Karakteristikat morfologjike për emrin

Pjesa e ligjëratës	***Etiketa PL***	***Karakteristikat morfologjike***	
		Shqip	**Etiketa**
Emri	NOUN	**Rasa**	**Case**
		Emërore	Nom
		Gjinore	Gen
		Dhanore	Dat
		Kallëzore	Acc
		Rrjedhore	Abl
		Trajta	**Definite**
		E shquar	Def
		E pashquar	Ind
		Gjinia	**Gender**
		Femërore	Fem
		Mashkullore	Masc
		Numri	**Number**
		Njëjës	Sing
		Shumës	Plur

Në gramatikën e gjuhës shqipe emri ka kategoritë gramatikore të rasës, trajtës, gjinisë dhe numrit. Çdo emër i përgjithshëm është etiketuar duke përcaktuar vlera për secilën nga këto kategori. Dallim është bërë tek emrat e përveçëm, ku emrat e frymorëve janë etiketuar duke përcaktuar vlera për kategorinë gramatikore të rasës dhe trajtës, kurse emrat e përveçëm të cilët mund të jenë togfjalësh janë përcaktuar karakteristika sipas pjesëve të ligjëratës që i përket secila fjalë e togfjalëshit.

Për shembull, nëse kemi emrin "Prenga" atëherë etiketimi është si në Tabelën 6.27

Tabela 6.27 Shembull i etiketimit të emrit të përveçëm të një frymori

Fjala	***Tema***	***Etiketa PL***	***Karakteristikat morfologjike***
Prenga	Prenga	PROPN	Case=Nom|Definite=Def

Për shembull, nëse kemi togfjalëshin "në Liceun Francez" atëherë etiketimi është si në Tabelën 6.28.

Tabela 6.28 Shembull i etiketimit të një emri të përveçëm të një jofrymori

Fjala	***Tema***	***Etiketa PL***	***Karakteristikat morfologjike***
në	në	ADP	_
Liceun	lice	PROPN	Case=Acc|Definite=Def|Gender=Masc|Number=Sing
Francez	francez	PROPN	Case=Acc|Degree=Pos|Gender=Masc|Number=Sing

Mbiemri në gramatikën e gjuhës shqipe ka kategoritë gramatikore të gjinisë, të numrit dhe rasës dhe i përshtatet në këto kategori emrit që ai përcakton. Mbiemrat ndahen në të nyjshëm dhe të panyjshëm. Skema e përdorur etiketon në nivel fjale dhe në rastin tonë nuk mund të përcaktojmë nëse një mbiemër është i nyjshëm ose jo. Në dallim nga emri që përcakton, mbiemri ka dhe kategorinë e shkallës. Ne kemi përcaktuar vlerë për kategorinë e rasës, gjinisë dhe numrit. Për kategorinë e shkallës, kemi përcaktuar vetëm shkallën pohore. Shkalla krahasore dhe sipërore është e pamundur të përcaktohet në këtë skemë duke qenë se janë forma analitike të krijuara nga mjete leksikore. Në Tabelën 6.29 janë specifikuar karakteristikat morfologjike të përdorura për etiketimin e mbiemrit.

Përemri në gjuhën shqipe klasifikohet në shtatë lloje siç kemi shpjeguar në çështjen 4.1.4 dhe në varësi të llojit mund të ketë kategorinë gramatikore të rasës, gjinisë, numrit dhe vetës. Për të gjitha llojet e përemrave është përdorur etiketa e pjesëve të ligjëratës, PRON, dhe për të përcaktuar llojin e tij është përdorur etiketa e karakteristikës gramatikore PronType. Në skemën e përdorur, përemri pronor dhe vetvetor ka të njëjtën vlerë të karakteristikës gramatikore PronType=Prs, kështu për identifikuar secilin nga këto përemra, në rastin e përemrit pronor është përdorur etiketa e kategorisë gramatikore Poss=Yes, kurse në rastin e përemrit vetvetor është përdorur etiketa e kategorisë gramatikore Reflex=Yes.

Tabela 6.29 Karakteristikat morfologjike për mbiemrin

Pjesa e ligjëratës	***Etiketa PL***	***Karakteristikat morfologjike***	
		Shqip	**Etiketa**
Mbiemri	ADJ	**Rasa**	**Case**
		Emërore	Nom
		Gjinore	Gen
		Dhanore	Dat
		Kallëzore	Acc
		Rrjedhore	Abl
		Gjinia	**Gender**
		Femërore	Fem
		Mashkullore	Masc
		Numri	**Number**
		Njëjës	Sing
		Shumës	Plur

Tabela 6.30 Karakteristikat morfologjike për përemrin

Pjesa e ligjëratës	***Etiketa PL***	***Karakteristikat morfologjike***		
		Shqip	**Etiketa**	
Përemri	PRON	**Rasa**	**Case**	
		Emërore	Nom	
		Gjinore	Gen	
		Dhanore	Dat	
		Kallëzore	Acc	
		Rrjedhore	Abl	
		Gjinia	**Gender**	
		Femërore	Fem	
		Mashkullore	Masc	
		Numri	**Number**	
		Njëjës	Sing	
		Shumës	Plur	
		Veta	**Person**	1, 2, 3
		Tipi	**PronType**	
		Pronorë	Prs	**Poss=Yes**
		Pyetës	Int	
		Dëftor	Dem	
		Vetorë	Prs	
		Lidhor	Rel	
		Vetvetor	Prs	**Reflex=Yes**
		Të papërcaktuar	Ind	

Në Tabelën 6.30 janë specifikuar karakteristikat morfologjike të përdorura për etiketimin e përemrit.

Në gjuhën shqipe ka pesë lloje ndajfoljesh. Për të përcaktuar tipin e ndajfoljes është përdorur karakteristika morfologjike AdvType. Ndajfoljet ashtu si mbiemrat kanë kategorinë gramatikore të shkallës. Edhe në këtë rast shkallët e ndryshme formohen nga forma analitike të krijuara nga mjete leksikore dhe nuk mund të etiketohet në skemën tonë, që bazohet në etiketim në nivel fjale. Në Tabelën 6.31 janë specifikuar karakteristikat morfologjike të përdorura për etiketimin e ndajfoljes.

Tabela 6.31 Karakteristikat morfologjike për ndajfoljen

Pjesa e ligjëratës	***Etiketa PL***	***Karakteristikat morfologjike***	
		Shqip	**Etiketa**
Ndajfolja	ADV	**Lloji**	**AdvType**
		Mënyrës	Man
		Sasisë	Deg
		Kohës	Tim
		Vendit	Loc
		Shkakut	Cau

Nyja është etiketuar me etiketën e pjesëve të ligjëratës DET, dhe në varësi se çfarë tipi nyje është janë përcaktuar karakteristikat morfologjike të përdorura. Nëse një nyje është nyje e rasës gjinore atëherë është përdorur në etiketim vetëm rasa. Nëse një nyje është nyje e mbiemrit atëherë në etiketim janë përdorur karakteristikat gramatikore të rasës, gjinisë, numrit dhe tipit me vlerë PronType=Art. Në Tabelën 6.32 janë specifikuar karakteristikat morfologjike të përdorura për etiketimin e nyjes.

Parafjala është pjesë e pandryshueshme e ligjëratës por që klasifikohet sipas rasës që përdoret. Në korpusin tonë parafjalët janë etiketuar me etiketën e karakteristikës morfologjike e rasës.

Edhe lidhëza është pjesë e pandryshueshme e ligjëratës dhe klasifikohen sipas funksionit sintaksor, në lidhëza bashkërenditëse dhe në lidhëza nënrenditëse. Etiketa CCONJ është përdorur për të etiketuar lidhëzat bashkërenditëse dhe etiketa SCONJ për të etiketuar lidhëzat nënrenditëse. Lidhëzat nuk janë etiketuar me karakteristika morfologjike.

Për numërorin është përdorur vetëm karakteristika morfologjike e tipit.

Pjesëza është pjesë e pandryshueshme e ligjëratës dhe nuk ka kategori gramatikore. Për këtë arsye nuk kemi përcaktuar në etiketim ndonjë kategori morfologjike.

Për pjesët e ligjëratës pasthirrmat, simbolet dhe shenjat e pikësimit nuk kemi përcaktuar karakteristika morfologjike.

Në Tabelën 6.33 janë specifikuar karakteristikat morfologjike të përdorura për etiketimin e lidhëzës, parafjalës, pjesëzës, pasthirrmës, simboleve, shenjave të pikësimit dhe numërorit.

Tabela 6.32 Karakteristikat morfologjike për nyjen

Pjesa e ligjëratës	***Etiketa PL***	***Karakteristikat morfologjike***	
		Shqip	**Etiketa**
Nyja	DET	**Rasa**	**Case**
		Emërore	Nom
		Gjinore	Gen
		Dhanore	Dat
		Kallëzore	Acc
		Rrjedhore	Abl
		Gjinia	**Gender**
		Femërore	Fem
		Mashkullore	Masc
		Numri	**Number**
		Njëjës	Sing
		Shumës	Plur
		Tipi	**PronType**
			Art

Tabela 6.33 Karakteristikat morfologjike për pjesët e tjera të ligjëratës

Pjesa e Ligjëratës	***Etiketa PL***	***Karakteristikat morfologjike***	
		Shqip	**Etiketa**
Lidhëza	SCONJ	Nënrenditëse	nuk ka
	CCONJ	Bashkërenditëse	nuk ka
Parafjala	ADP	Rasa	**Case**
Pjesëza	PART		nuk ka
Pasthirrma	INTJ		nuk ka
Simbole	SYM		nuk ka
Shenjat e pikësimit	PUNCT		nuk ka
Numërori	NUM	**Tipi**	**NumType**
			Card
			Ord

Të gjithë fjalët e huaja të përdorura në korpus janë etiketuar me etiketën e pjesëve të ligjëratës që i përkasin sipas rastit dhe me etiketën e karakteristikave morfologjike Foreign=Yes.

Për të identifikuar që midis një fjale dhe një shenje pikësimi nuk ka hapësirë është përdorur etiketa AfterSpace=No në fushën MISC të formatit CoNLL-U.

Etiketimi i foljeve përbën një rast specifik dhe sfidues. Duke u bazuar në karakteristikat e foljeve në gjuhën shqipe është përcaktuar dhe skema e etiketimit të foljeve. Në vijim përcaktohet në mënyrë të detajuar se si janë etiketuar të gjithë mënyrat dhe kohët e foljeve.

Në gjuhën shqipe kohët e foljeve ndahen në të thjeshta dhe në të përbëra. Kohët e përbëra formohen duke përdorur pjesëza, foljen ose pjesoren, të cilat janë etiketuar mbi bazën e zgjedhimit që ka secila nga pjesët duke patur parasysh që në princip në skemën që ne kemi përdorur çdo fjalë duhet të etiketohet individualisht. Në kohët e përbëra foljet kam ose jam janë etiketuar si folje ndihmëse. Kategoritë gramatikore të përdorura për foljen janë:

- Mënyra (Mood): për të përcaktuar mënyrën që i përket folja;
- Koha (Tense): për të përcaktuat nëse folja është kohën e tashme apo të shkuar;
- Aspekti (Aspect): për të përcaktuar një folje në kohën e pakryer ose në kohën e kryer të thjeshtë;
- Numri (Number): për të përcaktuar numrin njëjës ose shumës;
- Veta (Person): për të përcaktuar vetën e foljes;
- Diateza (Voice): për të përcaktuar zgjedhimin vepror apo jovepror.

Mënyra dëftore është mënyra që ka numrin më të madh të kohëve. Në këtë mënyrë kemi tri kohë të thjeshta dhe shtatë kohë të përbëra. Në Tabelën 6.34 tregohen mënyra se si janë etiketuar foljet në kohët e thjeshta të mënyrës dëftore.

Tabela 6.34 Etiketimi i kohëve të thjeshta të mënyrës dëftore

Mënyra Dëftore						
Koha	**Aspect**	**Mood**	**Number**	**Person**	**Tense**	**Voice**
E tashme	-	Ind	Sing/ Plur	1/2/3	Pres	Act/ Pass
E pakryera	Imp	Ind	Sing/ Plur	1/2/3	Pres	Act/ Pass
E kryera e thjeshtë	Perf	Ind	Sing/ Plur	1/2/3	Pres	Act/ Pass

Kohët e shkuara të përbëra të mënyrës dëftore krijohen duke përdorur foljen ndihmëse kam në formën veprore dhe jam në formë joveprore të zgjedhuara në një nga kohët: e tashme, e kryer ose e kryer e thjeshtë plus pjesorja e foljes. Kështu, në Tabelën 6.35 paraqitet mënyra e etiketimit të foljeve në këto kohë.

Tabela 6.35 Etiketimi i kohëve të shkuara të përbëra të mënyrës dëftore

Mënyra Dëftore

Koha		Aspect	Mood	Number	Person	Tense	Voice	Verb Form
E kryera	kam/jam	-	Ind	Sing/Plur	1/2/3	Pres	Act/Pass	
	pjesorja	-	-	-	-	-	-	Part
Me se e kryera	kam/jam	Imp	Ind	Sing/Plur	1/2/3	Past	Act/Pass	
	pjesorja	-	-	-	-	-	-	Part
E kryera e tejshkuar	kam/jam	Perf	Ind	Sing/Plur	1/2/3	Past	Act/Pass	
	pjesorja	-	-	-	-	-	-	Part

Kohët e ardhme të mënyrës dëftore krijohen duke përdorur pjesëzat do dhe të dhe foljen e zgjedhuar në një kohë të caktuar ose pjesoren. Pjesëza do dhe të janë etiketuar vetëm me etiketën e pjesëve të ligjëratës PART.

Koha e ardhme e mënyrës dëftore në diatezën veprore është etiketuar si në Tabelën 6.36 kurse në diatezën joveprore është etiketuar si në Tabelën 6.37.

Tabela 6.36 Etiketimi i kohës së ardhme të mënyrës dëftore diateza veprore

Mënyra Dëftore

Koha		Aspect	Mood	Number	Person	Tense	Voice
E ardhme	do						
	të						
	folja	-	Sub	Sing/Plur	1/2/2003	Pres	Act

Tabela 6.37 Etiketimi i kohës së ardhme të mënyrës dëftore diateza joveprore

Mënyra Dëftore

Koha		Aspect	Mood	Number	Person	Tense	Voice
E ardhme	do						
	të						
	folja	-	Ind	Sing/Plur	1/2/2003	Pres	Pass

Koha e ardhme e përparme e mënyrës dëftore formohet nga pjesëza do + foljen ndihmëse kam në diatezën veprore ose jam në diatezën joveprore në kohën e tashme të mënyrës lidhore + pjesore. Në Tabelën 6.38 tregohet mënyra e etiketimit të foljes në këtë kohë.

Tabela 6.38 Etiketimi i kohës së ardhme të përparme të mënyrës dëftore

Mënyra Dëftore

Koha		Aspect	Mood	Number	Person	Tense	Voice	Verb Form
E ardhme e përparme	do							
	të							
	kam/jam	-	Sub	Sing/Plur	1/2/3	Pres	Act/Pass	
	folja							Part

Koha e ardhme e përparme e së shkuarës e mënyrës dëftore formohet nga pjesëza do + foljen ndihmëse kam në diatezën veprore ose jam në diatezën joveprore në kohën e pakryer të mënyrës lidhore + pjesore. Koha e pakryer e mënyrës lidhore formohet nga pjesëza të + folja në kohën e pakryer të mënyrës dëftore. Në Tabelën 6.39 tregohet mënyra e etiketimit të foljes në këtë kohë.

Tabela 6.39 Etiketimi i kohës së ardhme të përparme të mënyrës dëftore

Mënyra Dëftore

Koha		Aspect	Mood	Number	Person	Tense	Voice	Verb Form
E ardhme e përparme e së shkuarës	do							
	të							
	kam/jam	Imp	Ind	Sing/ Plur	1/2/3	Pres	Act/ Pass	
	folja							Part

Siç kuptohet nga etiketimi i realizuar, në këtë skemë etiketimi nuk kemi përcaktimin dhe përdorimin e karakteristikës Tense=Fut që përcakton kohën e ardhme.

Në mënyrën lidhore kemi 4 kohë: dy kohë të thjeshta dhe dy kohë të përbëra. Kohët e thjeshta formohen nga pjesëza të + foljen në zgjedhimin përkatës. Kohët e përbëra formohen nga pjesëza të + folja kam për diatezën veprore ose jam për diatezën joveprore në zgjedhimin përkatës + pjesorja. Pjesëza të është etiketuar vetëm me etiketën e pjesëve të ligjëratës PART. Në Tabelën 6.40 tregohet mënyra se si janë etiketuar foljet në këtë mënyrë për secilën kohë.

Tabela 6.40 Etiketimi i kohëve të mënyrës lidhore

Mënyra Lidhore

Koha		Aspect	Mood	Number	Person	Tense	Voice	Verb Form
E tashme	të							
	folja	-	Sub	Sing/Plur	1/2/3	Pres	Act	
E tashme joveprore	të							
	folja	-	Ind	Sing/Plur	1/2/3	Pres	Pass	
E pakryer	të							
	folja	Imp	Ind	Sing/Plur	1/2/3	Past	Act/Pass	
E Kryer e thjeshtë	të							
	kam/jam		Sub	Sing/Plur	1/2/3	Pres	Act/Pass	
	pjesorja							Part
Më se e kryer	të							
	kam/jam	Imp	Ind	Sing/Plur	1/2/3	Past	Act/Pass	
	pjesorja							Part

Edhe në mënyrën habitore kemi 4 kohë: dy kohë të thjeshta dhe dy kohë të përbëra. Në zgjedhimin jovepror, në kohën e tashme dhe të pakryer folja ka të njëjtën formë me foljen në zgjedhimin vepror duke i shtuar para pjesëzën u. Duke qenë se skema që ne kemi përdorur bazohet në etiketimin në nivel fjale, në këtë rast nuk kemi etiketim të ndryshëm për diatezën joveprore. Në kohët e përbëra si në gjithë mënyrat e tjera dhe në këtë mënyre në zgjedhimin vepror përdoret folja jam kurse në atë jovepror folja kam.

Në Tabelën 6.41 tregohet se si është etiketuar folja në këtë mënyrë për secilën kohë.

Tabela 6.41 Etiketimi i kohëve të mënyrës habitore

Mënyra Habitore

Koha		Aspect	Mood	Number	Person	Tense	Voice	Verb Form
E tashme	folja	-	Adm	Sing/Plur	1/2/3	Pres	Act	
E tashme joveprore	u							
	folja	-	Adm	Sing/Plur	1/2/3	Pres	Act	
E Pakryer	folja	Imp	Adm	Sing/Plur	1/2/3	Past	Act	
E Pakryer joveprore	u							
	folja	Imp	Adm	Sing/Plur	1/2/3	Past	Act	
E Kryer e Thjeshtë	kam/jam	-	Adm	Sing/Plur	1/2/3	Pres	Act/Pass	
	pjesorja							Part
Më se e Kryer	kam/jam	Imp	Adm	Sing/Plur	1/2/3	Past	Act/Pass	
	pjesorja							Part

Mënyra dëshirore ka dy kohë: koha e tashme dhe koha e kryer. Në kohën e tashme zgjedhimi jovepror ka formën: pjesëza u + folja e zgjedhuar njëlloj si në zgjedhimin vepror. Prandaj, edhe në këtë rast nuk kemi etiketim të ndryshëm të zgjedhimit në diatezën veprore dhe joveprore. Në kohën e kryer në zgjedhimin vepror kemi përdorimin e foljes kam kurse në zgjedhimin jovepror të foljes jam. Në Tabelën 6.42 tregohet mënyra se si janë etiketuar foljet në këtë mënyrë për secilën kohë.

Tabela 6.42 Etiketimi i kohëve të mënyrës dëshirore

Mënyra Dëshirore

Koha		**Aspect**	**Mood**	**Number**	**Person**	**Tense**	**Voice**	**Verb Form**
E tashme	folja	-	Opt	Sing/Plur	1/2/3	Pres	Act	
E tashme (jo veprore)	u							
	folja	-	Opt	Sing/Plur	1/2/3	Pres	Act	
E Kryera	kam/jam	-	Opt	Sing/Plur	1/2/3	Pres	Act/Pass	
	pjesorja							Part

Edhe në mënyrën kushtore ashtu si në mënyrën dëshirore kemi vetëm dy kohë: koha e tashme dhe koha e kryer. Në Tabelën 6.43 tregohet mënyra se si janë etiketuar foljet në këtë mënyrë për secilën kohë.

Tabela 6.43 Etiketimi i kohëve të mënyrës kushtore

Mënyra Kushtore

Koha		**Aspect**	**Mood**	**Number**	**Person**	**Tense**	**Voice**	**Verb Form**
E Tashme	do							
	të							
	folja	Imp	Ind	Sing/Plur	1/2/3	Pass	Act/ Pass	
E Kryer	do							
	të							
	kam/jam	Imp	Ind	Sing/Plur	1/2/3	Pass	Act/ Pass	
	folja							Part

Në mënyrën urdhërore kemi vetëm një kohë, koha e tashme. Në Tabelën 6.44 tregohet mënyra se si është etiketuar folja në këtë kohë.

Tabela 6.44 Etiketimi i kohëve të mënyrës urdhërore

Mënyra Urdhërore						
Koha	**Aspect**	**Mood**	**Number**	**Person**	**Tense**	**Voice**
E folja Tashme	-	Imp	Sing/Plur	2	Pres	Act/Pass

Në gjuhën shqipe kemi përdorimin e katër formave të pashtjelluara foljore:

- Pjesorja;
- Forma e pashtjelluar mohore;
- Përcjellorja;
- Paskajorja.

Në vazhdim trajtohet mënyra se si janë etiketuar secila nga këto forma të pashtjelluara. Në Tabelën 6.45 tregohet mënyra e etiketimit të pjesores, kurse në Tabelën 6.46 tregohet një shembull.

Tabela 6.45 Etiketimi i pjesores së foljes

Forma	***Etiketa PL***	***Karakteristikat morfologjike***
Pjesore	VERB	VerbForm=Part

Tabela 6.46 Shembull i etiketimit të pjesores

Fjala	***Tema***	***Etiketa PL***	***Karakteristikat morfologjike***
mësuar	mësoj	VERB	VerbForm=Part

Në Tabelën 6.47 tregohet mënyra e etiketimit të formës së pashtjelluar mohore, kurse në Tabelën 6.48 dhe në Tabelën 6.49 tregohet nga një shembull etiketimi.

Tabela 6.47 Etiketimi i formës së pashtjelluar mohore

Diateza	***Forma***	***Etiketa PL***	***Karakteristikat morfologjike***
Veprore	pa	PART	
	pjesore	VERB	VerbForm=Part
Joveprore	pa	PART	
	u	PART	
	pjesore	VERB	VerbForm=Part

Tabela 6.48 Shembull etiketimi i formës së pashtjelluar mohore në diatezën veprore

Fjala	***Tema***	***Etiketa PL***	***Karakteristika morfologjike***
pa	pa	PART	
mësuar	mësoj	VERB	VerbForm=Part

Tabela 6.49 Shembull etiketimi i formës së pashtjelluar mohore në diatezën joveprore

Fjala	***Tema***	***Etiketa PL***	***Karakteristika morfologjike***
pa	pa	PART	
u	u	PART	
mësuar	mësoj	VERB	VerbForm=Part

Në Tabelën 6.50 tregohet mënyra e etiketimit të përcjellores, kurse në Tabelën 6.51 dhe në Tabelën 6.52 tregohet nga një shembull etiketimi.

Tabela 6.50 Etiketimi i përcjellores

Diateza	***Forma***	***Etiketa PL***	***Karakteristikat morfologjike***
Veprore	duke	PART	
	pjesore	VERB	VerbForm=Part
Joveprore	duke	PART	
	u	PART	
	pjesore	VERB	VerbForm=Part

Tabela 6.51 Shembull etiketimi i përcjellores në diatezën veprore

Fjala	***Tema***	***Etiketa PL***	***Karakteristikat morfologjike***
duke	duke	PART	
mësuar	mësoj	VERB	VerbForm=Part

Tabela 6.52 Shembull etiketimi i përcjellores në diatezën joveprore

Fjala	***Tema***	***Etiketa PL***	***Karakteristikat morfologjike***
duke	duke	PART	
u	u	PART	
mësuar	mësoj	VERB	VerbForm=Part

Në Tabelën 6.53 tregohet mënyra e etiketimit të paskajores, kurse në Tabelën 6.54 dhe në Tabelën 6.55 tregohet nga një shembull etiketimi.

Tabela 6.53 Etiketimi i paskajores

Diateza	***Forma***	***Etiketa PL***	***Karakteristikat morfologjike***
Veprore	për	PART	
	të	PART	
	pjesore	VERB	VerbForm=Part
Joveprore	për	PART	
	t'u		
	të	PART	
	u	PART	
	pjesore	VERB	VerbForm=Part

Tabela 6.54 Shembull etiketimi i paskajores në diatezën veprore

Fjala	***Tema***	***Etiketa PL***	***Karakteristikat morfologjike***
për	për	PART	
të	të	PART	
mësuar	mësoj	VERB	VerbForm=Part

Tabela 6.55 Shembull etiketimi i paskajores në diatezën joveprore

Fjala	***Tema***	***Etiketa PL***	***Karakteristikat morfologjike***
për	për	PART	
t’u			
të	të	PART	
u	u	PART	
mësuar	mësoj	VERB	VerbForm=Part

Në Figurën 6.56 tregohet shembulli i një fjalie të etiketuar.

Tabela 6.56 Shembull i etiketimit të një fjalie

 # sent_id = 351

# text =					
1	S’	s’	PART		SpaceAfter=No
2	kaloi	kaloj	VERB	Aspect=Perf|Mood=Ind|Number=Sing|Person=3|Tense=Past|Voice=Act	_
3	pak	pak	ADV	AdvType=Deg	_
4	dhe	dhe	CCONJ	_	_
5	nga	nga	ADP	Case=Nom	_
6	dheu	dhe	NOUN	Case=Nom|Definite=Def|Gender=Masc|Number=Sing	_
7	doli	dal	VERB	Aspect=Perf|Mood=Ind|Number=Sing|Person=3|Tense=Past|Voice=Act	_
8	një	një	DET	PronType=Art	_
9	rënkim	rënkim	NOUN	Case=Acc|Definite=Ind|Gender=Masc|Number=Sing	_
10	i	i	DET	Case=Acc|Gender=Masc|Number=Sing	_
11	çjerrë	çjerrë	ADJ	Case=Acc| Degree=Pos |Gender=Masc|Number=Sing	SpaceAfter=No
12	,	,	PUNCT	_	_
13	që	që	PRON	PronType=Rel	_
14	s'	s'	PART	_	SpaceAfter=No
15	do	do	PART	_	_
16-17	ta	_	_	_	_
16	të	të	PART	_	_
17	e	e	PRON	Case=Acc|Gender=Masc|Number=Sing|Person=3|PronType=Prs	_
18	harroj	harroj	VERB	Mood=Sub|Number=Sing|Person=1|Tense=Pres|Voice=Act	_
19	kurrë	kurrë	ADV	AdvType=Tim	SpaceAfter=No
20	.	.	PUNCT	_	_

6.2.3. Vlerësimi eksperimental

Parsuesi Turku Pipeline është përdorur për t'u trajnuar dhe vlerësuar duke përdorur korpusin e etiketuar të përshkruar në çështjen 6.2.1 sipas skemës së shpjeguar në çështjen 6.2.2. Në Figurën 6.8 është paraqitur skema e trajnimit dhe vlerësimit të modelit të etiketuesit të realizuar.

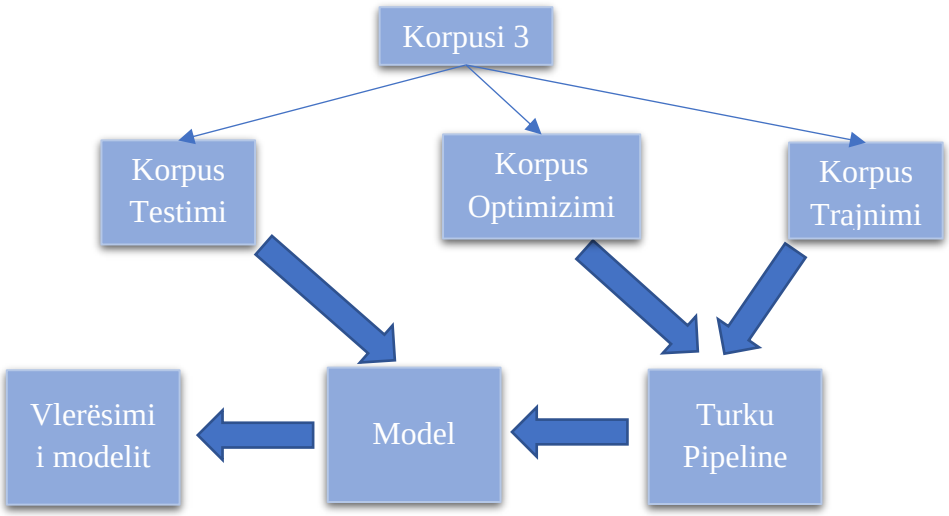


Figura 6.8 Skema e trajnimit dhe vlerësimit të modelit

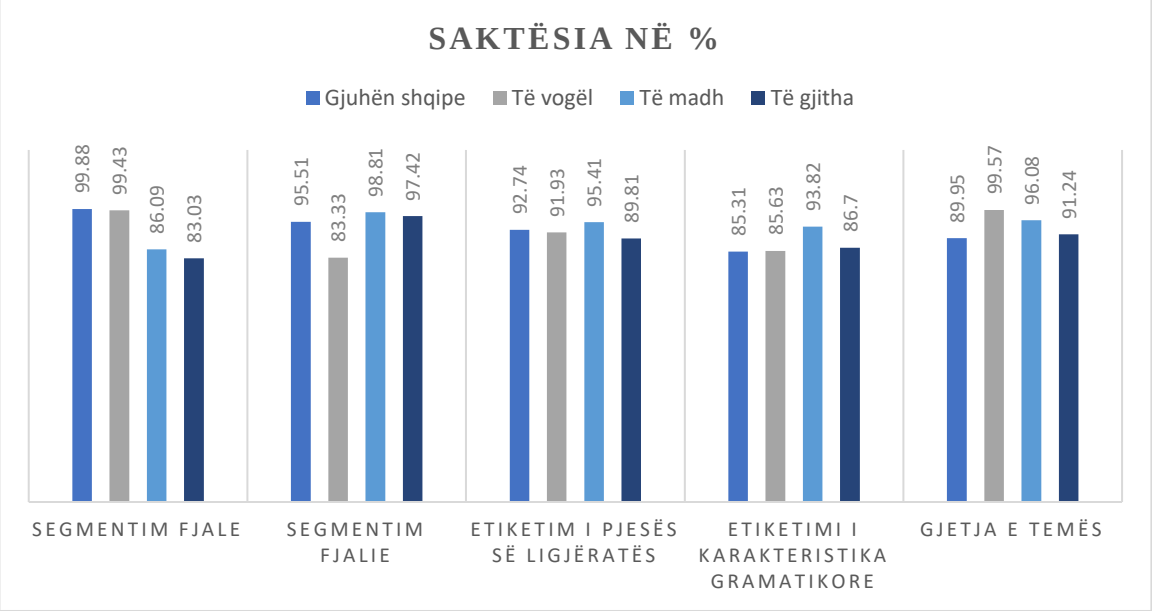
Në Tabelën 6.57 tregohen rezultatet e arritura në fazën e testimit duke përdorur skriptin vlerësues zyrtar nga CoNLL 2018 (Zeman, et al., 2018). Ne kemi vlerësuar saktësinë e etiketimit në përqindje për segmentimin në fjalë dhe fjali, për etiketimin e pjesëve të ligjëratës, për etiketimin e karakteristikave gramatikore dhe për gjetjen temës së fjalës.

Tabela 6.57 Rezultatet e vlerësimit të sistemit

	Saktësia [%]
Segmentim fjale	99.88
Segmentim fjalie	99.51
Etiketim i pjesëve të ligjëratës	92.74
Etiketimi i karakteristika gramatikore	85.31
Gjetja e temës	89.95

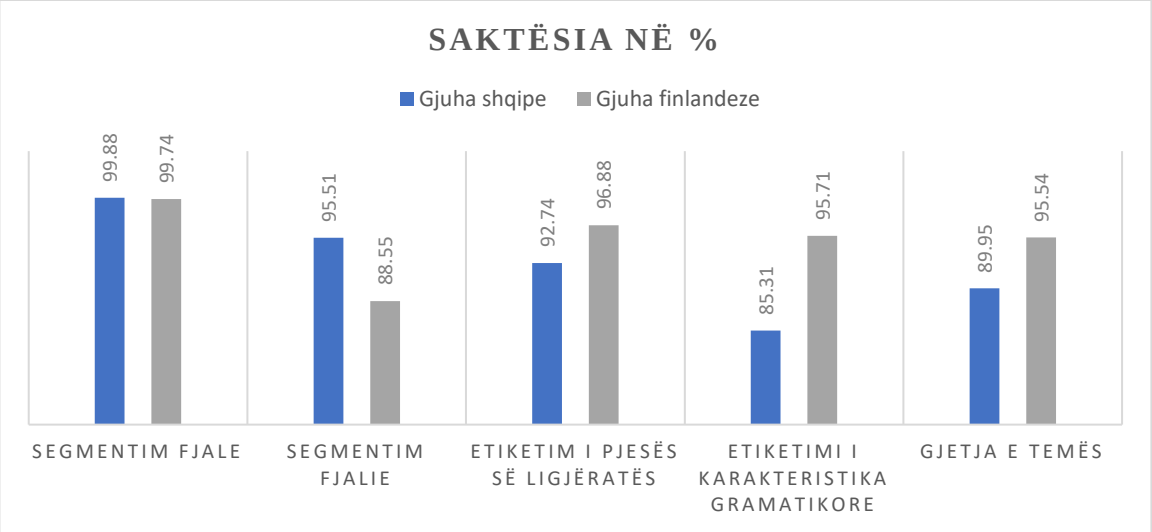
Nga rezultatet e arritura mund të themi që etiketuesi i propozuar në këtë disertacion ka performancë shumë të mirë në çdo fazë të tij, dhe është mjaftueshëm i mirë për t'u përdorur si ndihmë në aplikacione të tjera të përpunimit të gjuhës natyrale.

Në Grafikun 6.2 tregohet krahasimi i saktësisë së modelit të etiketuesit për gjuhën shqipe, me saktësinë mesatare të modeleve të trajnuara për gjuhët që kanë korpus të madh të dhënash të etiketuara, me saktësinë mesatare të modeleve të trajnuara për gjuhët që kanë korpus të vogël të dhënash të etiketuara dhe me saktësinë mesatare të të gjithë modeleve të trajnuara për të gjitha gjuhët e tjera përveç gjuhës shqipe



Grafiku 6.2 Saktësi e modeleve të ndryshme etiketuesish që përdorin Turku Pipeline

Në Grafikun 6.3 tregohet krahasimi i saktësisë të etiketuesit për gjuhën shqipe me saktësinë e etiketuesit për gjuhën finlandeze.



Grafiku 6.3 Saktësia e modelit të gjuhës shqipen dhe modelit të gjuhës finlandeze

KREU 7

PËRFUNDIME

7.1. Kontributi i këtij disertacioni

Ky disertacion kontribuon në sistemet inteligjente për përpunimin e informacionit tekst dhe gjetjen e informacionit në tekste në gjuhën shqipe. Konkretisht në këtë disertacion është realizuar:

- Krijimi i një korpusi të etiketuar opinionesh të mbledhura nga mediet *online* në gjuhën shqipe sipas polaritetit të ndjenjës së shprehur në to, në pozitive dhe negative.
- Analizimi i performacës së 50 algoritmeve MA të implementuar në platformën WEKA dhe të dy rrjeteve neurale, një model *bag-of-words* dhe një model CNN të implementuar në platformat Keras dhe TensorFlow për detyrën e klasifikimit të opinioneve sipas polaritetit të ndjenjës së shprehur në to. Është vlerësuar ndikimi i mjeteve të ndryshme, si komponenti linguisti, TF-IDF dhe n-gram në performancën e këtyre algoritmeve. Gjithashtu është vlerësuar ndikimi i madhësisë së korpusit të trajnimit dhe të temave të opinioneve në performancën e algoritmeve.
- Propozimi i skemës së parë të etiketimit për pjesët e ligjëratës dhe karakteristikave morfologjike për tekst në gjuhën shqipe, bazuar në skemën shumë gjuhësore Universal Dependencies (UD).
- Krijimi i korpusit të parë të etiketuar në gjuhën shqipe me 184,597 fjalë (23,686 fjali) të etiketuara me etiketën e pjesëve të ligjëratës, me etiketat e karakteristikave morfologjike dhe me temën e fjalës.
- Krijimi i etiketuesit të parë morfologjik dhe temëzuesit të parë për gjuhën shqipe me akses publik dhe me performancë të krahasueshme me gjuhë të tjera të nivelit të kompleksitetit të gjuhës shqipe.
- Eksperimentimi i të gjitha teknikave dhe modeleve të propozuara duke treguar efektshmërinë e tyre.

7.2. Puna në të ardhmen

Zgjidhjet e propozuara në këtë disertacion bazohen në përdorimin e teknikave të të mësuarit e automatizuar të kontrolluar. Në këto sisteme përveç përmirësimit të algoritmit të përdorur dhe karakteristikave të konfigurimit, në trajnim një rol të rëndësishëm luan dhe korpusi i etiketuar nga i cili këto algoritme mësojnë. Në këtë aspekt, ndikim të madh ka cilësia e etiketimit të korpusit dhe madhësia e korpusit. Sa më saktë të jetë etiketuar një korpus dhe sa më shumë të dhëna të ketë aq më shumë përmirësohet performanca e sistemit.

Si punë në të ardhmen ne propozojmë:

- Zgjerimin e korpusit të opinioneve;

- Rishikimin e skemës së etiketimit morfologjik dhe të karakteristikave morfologjike nën asistencën e gjuhëtarëve për të realizuar një skemë sa më konform rregullave të gjuhës shqipe;
- Rishikimi i korpusit të etiketuar për gabime të mundshme në etiketim;
- Identifikimi i kategorive të pjesëve të ligjëratës dhe karakteristikave morfologjike që kanë saktësinë më të ulët të parashikimit nga ana e etiketuesit;
- Zgjerimi i korpusit të etiketuar me pjesët e ligjëratës, karakteristikat morfologjike dhe temën duke patur parasysh kategoritë e etiketimit ku etiketuesi ka saktësi më të ulët;
- Zgjerimi i korpusit me fjali me kompleksitet më të lartë nga ana gjuhësore;
- Krijimi i një skeme të etiketimit sintaksor për tekst në gjuhën shqipe;
- Etiketimi sintaksor i korpusit;
- Trajnimi i një parseri të varësisë për gjuhën shqipe.

BIBLIOGRAFIA

Abad, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., & Citro, Z. (2015). TensorFlow: Large-scale machine learning on heterogeneous systems.

Agalliu, F., Angoni, E., Demiraj, S., Dhrimo, A., Hysa, E., Lafe, E., & Likaj, E. (2002). *Gramatika e gjuhës shqipe, V1: Morfologjia.* Tiranë: Akademia e Shkencave.

AL-Sharuee, M., Liu , F., & Pratama, M. (2018). Sentiment analysis: An automatic contextual analysis and ensemble clustering approach and comparison. *Data & Knowledge Engineering*, 194-213.

Ange, T., Roger, N., Aude, D., & Claude, F. (2018). Semi-Supervised Multimodal Deep Learning Model for Polarity Detection in Arguments. *2018 International Joint Conference on Neural Networks (IJCNN)*, (fv. 1-8).

Arkhangelskıj, T., Danıel, M., Morozova, M., & Rusakov, A. (2011). Korpusı i Gjuhës Shqıpe: Drejtımet Kryesore të Punës. *Albanıan And Balkan Languages, Scientific Conference* (fv. 635-683). Prishtinë: ASHAK.

Arkhangelskiy, T., Belyaev, O., & Vydrin, A. (2012). The creation of large-scaled annotated corpora of minority languages using UniParser and the EANC platform. *Proceedings of COLING 2012: Posters*, (fv. 83–92). Mumbai, India.

Arunachalam , R., & Sarkar, S. (2013). The new eye of government: Citizen sentiment analysis in social media. *IJCNLP 2013 Workshop on Natural Language Processing for Social Media (SocialNLP)*, (fv. 23–28). Japan.

Asch, V., & Daelemans, W. (2016). Predicting the Effectiveness of Self-Training: Application to Sentiment Classification. *ArXiv*.

Bergmanis, T., & Goldwater, S. (2018). Context Sensitive Neural Lemmatization with Lematus. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)* (fv. 1391–1400). New Orleans, Louisiana: Association for Computational Linguistics.

Biba, M., & Gjati, E. (2014). Boosting Text Classification through Stemming of Composite Words. *Recent Advances in Intelligent Informatics, Advances in Intelligent Systems and Computing*, 185-194.

Brants, T. (2000). TnT — A Statistical Part-of-Speech Tagger. *ANLC '00: Proceedings of the sixth conference on Applied natural language processing*, (fv. 224–231).

BrightLocal. (2021). Gjetur në https://www.brightlocal.com/research/local-consumer-review-survey/.

Brill, E. (1992). A Simple Rule-Based Part of Speech Tagger. *ANLC '92: Proceedings of the third conference on Applied natural language processing*, (fv. 152–155).

Brill, E. (1995). Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study in Part-of-Speech Tagging. *Computational Linguistics, Volume 21, Number 4*, 543–565.

Brill, E., & Mooney, R. (1997). An Overview of Empirical Natural Language Processing. *AI Magazine, 18(4)*.

Buchholz, S., & Marsi, E. (2006). CoNLL-X shared task on Multilingual Dependency Parsing. *Proceedings of the 10th Conference on Computational Natural Language Learning (CoNLL-X)* (fv. 149–164). New York City: Association for Computational Linguistics.

Carter, D., & Inkpen, D. (2015). Inferring aspect-specific opinion structure in product reviews using co-training. *CICLing 2015: Computational Linguistics and Intelligent Text Processing*, (fv. 225–240).

Catal, C., & Nangir, M. (2017). A Sentiment Classification Model Based on Multiple Classifiers. *Applied Soft Computing Journal*, 135-141.

Çeliku, M., Domi, M., Floqi, S., Mansaku, S., Përnaska, R., Prifti, S., & Totoni, M. (2002). *Gramatika e gjuhës shqipe,V.2: Sintaksa.* Tiranë: Akademia e Shkencave.

Cessie, L., & Houwelingen, C. (1992). Ridge Estimators in Logistic Regression. *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, 191-201.

Chiavetta, F., Bosco, G., & Pilato, G. (2016). A Lexicon-based Approach for Sentiment Classification of Amazon Books Reviews in Italian Language. *12th International Conference on Web Information Systems and Technologies*, (fv. 159-170).

Choi, Y., & Lee, H. (2017). Data properties and the performance of sentiment classification. *Information Systems Frontiers*, 993–1012.

Chollet, F. (2015). *Keras.* Gjetur në https://github.com/fchollet/keras.

Christodoulopoulos, C., Goldwater, S., & Steedman, M. (2010). Two Decades of Unsupervised POS Induction: How Far Have We Come? *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing* (fv. 575-584). Cambridge, MA: Association for Computational Linguistics.

Clark, A., Fox, C., & Lappin, S. (2010). *The Handbook of Computational Linguistics and Natural Language Processing.* Blackwell Publishing Ltd.

Dale, R. (2010). Classical Approaches to Natural Language Processing. Në N. Indurkhya, & F. J. Damerau, *HANDBOOK OF NATURAL LANGUAGE PROCESSING* (fv. 3-8). Chapman & Hall/CRC.

Dale, R. (2014). Spoken Language Dialogue Systems. Në R. Mitkov, *The Oxford Handbook of Computational Linguistics 2nd edition.* Oxford University Press.

Dawson, J. (1974). Suffix removal and word conflation. *ALLC Bulletin*, 33-46.

Dinu, L., & Iuga, I. (2012). The Naive Bayes Classifier in Opinion Mining:In Search of the Best Feature Set. *International Conference on Intelligent Text Processing and Computational Linguistics. Computational Linguistics and Intelligent Text Processing. CICLing 2012.* (fv. 556-567). Berlin: Springer.

Dozat, T., Qi, P., & Manning, C. (2017). Stanford's Graph-based Neural Dependency Parser at the CoNLL 2017 Shared Task. *Proceedings of the CoNLL 2017 Shared*

Task: Multilingual Parsing from Raw Text to Universal Dependencies (fv. 20-30). Vancouver, Canada: Association for Computational Linguistics.

Federici, M., & Dragoni, M. (2017). A Branching Strategy For Unsupervised Aspect-based Sentiment Analysis. *Proceedings of the 3rd International Workshop on Emotions, Modality, Sentiment Analysis and the Semantic Web co-located with 14th ESWC 2017.*

Fojhtvanger, L. (1999). *Çifutka e Toledos.* Shtëpia botuese Dituria.

Gautam, G., & Yadav, D. (2014). Sentiment analysis of twitter data using machine learning approaches and semantic analysis. *2014 Seventh International Conference on Contemporary Computing (IC3)*, (fv. 437-442).

Genkin, A., Lewis, D., & Madigan, D. (2007). Large-Scale Bayesian Logistic Regression for Text Categorization. *Technometrics*, 291-304.

Goldhahn, D., Eckart, T., & Quasthoff, U. (2012). Building Large Monolingual Dictionaries at the Leipzig Corpora Collection: From 100 to 200 Languages. *The 8th International Conference on Language Resources and Evaluation.*

Grishman, R. (2014). Information Extraction. Në R. Mitkov, *The Oxford Handbook of Computational Linguistics 2nd edition.* Oxford University Press.

Gutiérrez, L., Bekios-Calfa, J., & Keith, B. (2018). A Review on Bayesian Networks for Sentiment Analysis. *International Conference on Software Process Improvement: Trends and Applications in Software Engineering* (fv. 111-120). Springer.

Habernal, I., Ptácek, T., & Steinberger, J. (2014). Supervised sentiment analysis in Czech social media. *Information Processing and Management*, 693-707.

Hajmohammadi , M., Ibrahim, R., & Selamat, A. (2014). Cross lingual sentiment classification using multiple source languages in multi-view semi-supervised learning. *Engineering Applications of Artificial Intelligence*, 195-203.

Hajmohammadi, M., Ibrahim, R., & Selamat, A. (2015). Graph-Based Semi-supervised Learning for Cross-Lingual Sentiment Classification. *ACIIDS 2015: Intelligent Information and Database Systems*, (fv. 97-106).

Harris, Z. (1962). *String Analysis of Sentence Structure.* MOUTON & CO . THE HAGUE .

Hong, S., Lee, J., & Lee, J. (2014). Competitive self-training technique for sentiment analysis in mass social media. *2014 Joint 7th International Conference on Soft Computing and Intelligent Systems (SCIS) and 15th International Symposium on Advanced Intelligent Systems (ISIS)*, (fv. 9-12).

Hovy, E. (2014). Text Summarization. Në R. Mitkov, *The Oxford Handbook of Computational Linguistics 2nd edition.* Oxford University Press.

Hu, M., & Liu, B. (2004). Mining and summarizing customer reviews. *KDD '04: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, (fv. 168–177).

Iosifidis, V., & Ntutsi, E. (2017). Large scale sentiment learning with limited labels. *KDD '17: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, (fv. 1823–1832).

İşgüder-Şahin, G., Zafer, H., & Adah, E. (2014). Polarity detection of Turkish comments on technology companies. *International Conference on Asian Language Processing (IALP)*, (fv. 136-139).

Jeyapriya, A., & Selvi, K. (2015). Extracting aspects and mining opinions in product reviews using supervised learning algorithm. *2015 2nd International Conference on Electronics and Communication Systems (ICECS)*, (fv. 548–552).

Joshi, A. K., & Hopely, P. (1996). A Parser from Antiquity. *Natural Language Engineering*, fv. 291–294.

Jurafsky, D., & Martin, J. (2020). *Speech and Language Processing (3rd ed. draft).*

Kabashi, B., & Proisl, T. (2016). A Proposal for a Part-of-Speech Tagset for the Albanian Language. *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)* (fv. 4305–4310). Portorož, Slovenia: European Language Resources Association (ELRA).

Kabashi, B., & Proisl, T. (2018). Albanian Part-of-Speech Tagging: Gold Standard and Evaluation. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)* (fv. 2593–2599). Miyazaki, Japan: European Language Resources Association (ELRA).

Kadare, I. (2011). *Kohë e pamjaftueshme.* Shtëpia botuese Onufri.

Kadriu, A. (2013). NLTK Tagger for Albanian using Iterative Approach. *Proceedings of the 35th International Conference on Information Technology Interfaces*, (fv. 283-288).

Kanerva, J., Ginter, F., & Salakoski, T. (2020). Universal Lemmatizer: A sequence-to-sequence model for lemmatizing Universal Dependencies treebanks. *Natural Language Engineering*, 1 - 30.

Kanerva, J., Ginter, F., Miekka, N., Leino, A., & Salakoski, T. (2018). Turku Neural Parser Pipeline: An End-to-End System for the CoNLL. *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies* (fv. 133–142). Brussels, Belgium: Association for Computational Linguistics.

Karanikolas, N. (2009). Bootstrapping the Albanian Information Retrieval. *Fourth Balkan Conference in Informatics*, (fv. 231-235).

Karanikolas, N. (2013). A methodology for building simple but robust stemmers without language knowledge – Overview, data model and ranking algorithm. *CompSysTech '13: Proceedings of the 14th International Conference on Computer Systems and Technologies* (fv. 284-290). Ruse, Bulgaria: ACM PRESS.

Karanikolas, N. (2014). A Methodology for Building Simple but Robust Stemmers without Language Knowledge: Stemmer Configuration. *Procedia - Social and Behavioral Sciences*, 370 – 375.

Kauffmanna, E., Peralb, J., Gilc, D., Ferrández, A., Sellers, R., & Mora, H. (2020). A framework for big data analytics in commercial social networks: A case study on sentiment analysis and fake review detection for marketing decision-making. *Industrial Marketing Management*, 523-537.

Khoo, C., & Johnkhan, S. (2018). Lexicon-based sentiment analysis: Comparative evaluation of six sentiment lexicons. *Journal of Information Science*, 491-511.

Kirilenko, A., Stepchenkova, S., Kim, H., & Li, X. (2018). Automated Sentiment Analysis in Tourism: Comparison of Approaches. *Journal of Travel Research*, 1012-1025.

Kirov, C., Cotterell, R., Sylak-Glassman, J., Walther, G., Vylomova, E., Xia, P., . . . Hulden, M. (2018). UniMorph 2.0: Universal Morphology. *The Eleventh International Conference on Language Resources and Evaluation.*

Klein, G., Kim, Y., Deng, Y., Senellart, J., & Rush, A. (2017). OpenNMT: open-source toolkit for neural machine translation. *Proceedings of ACL 2017, System Demonstrations* (fv. 67–72). Vancouver,Canada: Association for Computational Linguistics.

Kostallari, A., Samara, M., Kole, J., Daka, P., Haxhillazi, P., Shehu, H., . . . Hidi, A. (1984). *Fjalori i shqipes së sotme.* Akademia e Shkencave të Shqipërisë.

Kote, N., & Biba, M. (2018). An Experimental Evaluation of Algorithms for Opinion Mining in Multi-Domain Corpus in Albanian. *International Symposium on Methodologies for Intelligent Systems* (fv. 439-447). Springer, Cham.

Kote, N., & Biba, M. (2019). Opinion Mining: Analizimi i Teknikave që Përdoren për Klasifikimin e Opinioneve. *Buletini i Shkencave Teknike*, 47-61.

Kote, N., & Biba, M. (2020). Opinion Mining in Albanian: Evaluation of the Performance of Machine Learning Algorithms for Opinions Classification. *International Journal of Innovative Science and Research Technology*.

Kote, N., Biba, M., & Trandafili, E. (2018). A Thorough Experimental Evaluation of Algorithms for Opinion Mining in Albanian. *International Conference on Emerging Internetworking, Data & Web Technologies* (fv. 525-536). Springer, Cham.

Lafe, V., Buxheli, L., & Basha, N. (1979). *Libri i gjuhës shqipe 1.* Tiranë: Shtëpia botuese e librit shkollor.

Landwehr, N., Hall, M., & Frank, E. (2005). Logistic Model Trees. *Machine Learning*, 161–205.

Le, Q., & Mikolov, T. (2014). Distributed Representations of Sentences and Documents. *Proceedings of the 31st International Conference on Machine Learning, PMLR*, (fv. 1188-1196).

Li, G., & Liu, F. (2012). Application of a clustering method on sentiment analysis. *Journal of Information Science*, 127–139.

Li, Z., Zhao, H., & Parnow, K. (2020). Global Greedy Dependency Parsing. *The Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI-20)*, (fv. 8319-8326). New York, NY, USA.

Liddy, E. (1996, April/May). Enhanced text retrieval using natural language. *Bulletin of the American Society for Information Science and Technology*, fv. 14-16.

Liu, B. (2010). Sentiment Analysis and Subjectivity. Në N. Indurkhya, & F. Damerau, *Handbook of Natural Language Processing.* Chapman and Hall/CRC.

Liu, B. (2015). *Sentiment Analysis: Mining Opinions, Sentiments, and Emotions.* Cambridge University Press.

Lovins, J. (1968). Development of a Stemming Algorithm. *Mechanical Translation and Computational Linguistics*, 22-31.

Manning, C. D., & Schutze, H. (2000). *Foundations of statistical natural language processing.* Cambridge, Massachusetts;London, England: The MIT Press.

Mayfield, J., & McNamee, P. (2003). Single n-gram stemming. *SIGIR '03: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, (fv. 415-416).

McDonald, R., & Nivre, J. (March 2011). Analyzing and Integrating Dependency Parsers. *Computational Linguistics, Volume 37, Issue 1*, 197–230.

Mei, Q., & Radev, D. (2014). Information Retrieval. Në R. Mitkov, *The Oxford Handbook of Computational Linguistics 2nd edition.* Oxford University Press.

Mikolov, T., Grave, E., Bojanowski, P., Puhrsch, C., & Joulin, A. (2018). Advances in Pre-Training Distributed Word Representations. *Proceedings of the International Conference on Language Resources and Evaluation.* Gjetur në https://fasttext.cc/docs/en/crawl-vectors.html

Molina-González, M., Martínez-Cámara, E., Martín-Valdivia, M., & Ureña-López, L. (2015). A Spanish semantic orientation approach to domain adaptation for polarity classification. IPM, 51(4): 520-531, 2015. *Information Processing and Management: an International Journal*, 520–531.

Moraes, R., Valiati, J., & Neto, W. (2013). Document-level sentiment classification: an empirical comparison between SVM and ANN. *Expert Systems with Applications*, 621-633.

Morozova, M., & Rusakov, A. (2014). Albanian National Corpus: Composition, Text Processing and Corpus-oriented Grammar Development, Language and culture of the Albanian. *The 5th Deutsch-Albanischen Cultural Science Conference* (f. 270). Pogradec, Albania: Hubert & Co., Göttingen.

Nivre, J., de Marneffe, M., Ginter, F., Hajič, J., Manning, C., Pyysalo, S., . . . Zeman, D. (2020). Universal Dependencies v2: An Evergrowing Multilingual Treebank Collection. *Proceedings of the 12th Language Resources and Evaluation Conference* (fv. 4034–4043). Marseille, France: European Language Resources Association.

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. Në *Foundations and Trends in Information Retrieval* (fv. 1-135).

Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up?: sentiment classification using machine learning techniques. *Conference on Empirical Methods in Natural Language Processing* (fv. 79–86). Association for Computational Linguistics.

Park , S., & Kim, Y. (2016). Building Thesaurus Lexicon using Dictionary Based Approach for Sentiment Classification. *2016 IEEE 14th International Conference on Software Engineering Research, Management and Applications (SERA)*, (fv. 39-44).

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., & Grisel, O. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 2825-2830.

Pham, D., & Le, A. (2017). Learning multiple layers of knowledge representation for aspect based sentiment analysis. *Data & Knowledge Engineering*, 26-39.

Piton, O., & Lagji, K. (2007). Morphological study of Albanian words, and processing with NooJ. *NooJ Conference*, (fv. 189-205).

Piton, O., Lagji, K., & Përnaska, R. (2007). Electronic Dictionaries and Transducers for Automatic Processing of the Albanian Language. *12th International Conference on Applications of Natural Language to Information Systems*, (fv. 407-413).

Platt, J. (1999). Fast Training of Support Vector Machines using Sequential Minimal Optimization. Në *Advances in kernel methods: support vector learning* (fv. 185–208). MIT Press.

Poria, A., Cambria, E., & Gelbukh, A. (2016). Aspect extraction for opinion mining with a deep convolutional neural network. *Knowledge-Based Systems*, 42-49.

Porter, M. (1980). An algorithm for suffix stripping. *Electronic Library and Information Systems*, 211-218.

Porter, M. (2001). *Snowball: a language for stemming algorithms.*

Prager, J. (2014). Question Answering. Në R. Mitkov, *The Oxford Handbook of Computational Linguistics 2nd edition.* Oxford University Press.

Qi, P., Dozat, T., Zhang, Y., & Manning, C. D. (2018). Universal Dependency Parsing from Scratch. *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies* (fv. 160-170). Brussels, Belgium: Association for Computational Linguistics.

Ren, F., & Kang, X. (2013). Employing hierarchical Bayesian networks in simple and complex. *Computer Speech and Language*, 943–968.

Rennie, D., Shih, L., Teevan, J., & Karger, R. (2003). Tackling the Poor Assumptions of Naive Bayes Text Classifiers. In: ICML. *International Conference on Machine Learning*, (fv. 616-623). Washington DC.

Russell, S., & Norvig, P. (2020). *Artificial Intelligence A Modern Approach Book.* Pearson.

Sadiku, J., & Biba, M. (2012). Automatic Stemming of Albanian Through a Rule-based Approach. *Journal of International, Research Publications: Language, Individuals and Society*, 173-190.

Salavaçi, E., & Biba, M. (2012). Enhancing Part-of-Speech Tagging in Albanian with Large Tagsets.

Saleh, M., Martín-Valdivia, M., Montejo-RáezL, A., & Ureña-López, A. (2011). Experiments with SVM to classify opinions in different domains. *Expert Systems with Applications*, 14799-14804.

Severyn, A., Moschitti, A., Uryupina, O., Plank, B., & Filippova, K. (2014). Opinion Mining on YouTube. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics* (fv. 1252–1261). Baltimore, Maryland, USA: Association for Computational Linguistics.

Severyn, A., Moschitti, A., Uryupina, O., Plank, B., & Filippova, K. (2015). Multi-lingual opinion mining on YouTube. *Information Processing and Management*, 46-60.

Silva, F., Coletta, F., & Hruschka, E. (2016). A Survey and Comparative Study of Tweet Sentiment Analysis via Semi-Supervised. *ACM Computing Surveys*, 1–26.

Specia, L., & Wilks, Y. (2014). Machine Translation. Në R. Mitkov, *The Oxford Handbook of Computational Linguistics 2nd edition.* Oxford University Press.

Straka, M., & Strakova, J. (2017). Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe. *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies* (fv. 88–99). Vancouver, Canada: Association for Computational Linguistics.

Sumner, M., Frank, E., & Hall, M. (2005). Speeding Up Logistic Model Tree Induction. *Knowledge Discovery in Databases: PKDD 2005* (fv. 675–683). Springer.

Tang, D., Qin, B., Liu, T., & Yang, Y. (2015). User modeling with neural network for review rating prediction. *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence (IJCAI 2015)*, (fv. 1340–1346).

Teng, Z., Vo, D., & Zhang, Y. (2016). Context-Sensitive Lexicon Features for Neural Sentiment Analysis. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, (fv. 1629–1638).

Thelwall, M. (2019). Sentiment Analysis for Tourism. Në M. Sigala, R. Rahimi, & M. Thelwall, *Big Data and Innovation in Tourism, Travel, and Hospitality* (fv. 87-104). Springer.

Ting, K. (2011). Precision and Recall. Në C. Sammut, & G. Webb, *Encyclopedia of Machine Learning.* Boston, MA: Springer.

Toska, M., Nivre, J., & Zeman, D. (2020). Universal Dependencies for Albanian. *Proceedings of the Fourth Workshop on Universal Dependencies (UDW 2020)* (fv. 178–188). Barcelona, Spain (Online): Association for Computational Linguistics.

Toutanova, K., Klein, D., Manning, C., & Singer, Y. (2003). Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network. *In Proceedings of HLT-NAACL 2003*, (fv. 252-259).

Tripathy, A., Agrawal , A., & Rath, S. (2016). Classification of sentiment reviews using n-gram machine learning approach. *Expert Systems with Applications*, 117-126.

Trommer, J., & Kallulli, D. (2004). A Morphological Tagger for Standard Albanian. *In Proceedings of LREC 2004.* Lisbon, Portugal.

Tsai, C. F., Chen, K., Hu, Y. H., & Chen, W. K. (2020). Improving text summarization of online hotel reviews with review helpfulness and sentiment. *Tourism Management*.

Tsakalidis, A., Papadopoulos, S., Cristea, A., & Kompatsiaris, Y. (2015). Predicting elections for multiple countries using Twitter and polls. *IEEE Intelligent Systems*, 10-17.

Tufiș, D., & Ion, I. (2014). Part-of-Speech Tagging. Në R. Mitkov, *The Oxford Handbook of Computational Linguistics 2nd edition.* Oxford University Press.

Unnisa, M., Ameen, A., & Raziuddin, S. (2016). Opinion Mining on Twitter Data using Unsupervised Learning Technique. *International Journal of Computer Applications*, 12-19.

Uryupina, O., Plank, B., Severyn, A., & Moschitti, A. (2014). Media, SenTube: A Corpus for Sentiment Analysis on YouTube Social. *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)* (fv. 4244–4249). European Language Resources Association (ELRA).

Vilares, D., Gomez-Rodriguez, C., & Alonso, M. (2017). Universal, Unsupervised (Rule-Based), Uncovered Sentiment Analysis. *Knowledge-Based Systems*, 45-55.

Voutilainen, A. (2005). Part-of-Speech Tagging. Në R. Mitkov, *The Oxford Handbook of Computational Linguistics (1 ed.).* Oxford University Press.

Wan, Y., & Gao Q., Q. (2015). An Ensemble Sentiment Classification System of Twitter Data for Airline Services Analysis. *2015 IEEE International Conference on Data Mining Workshop (ICDMW)*, (fv. 1318-1325).

Wang, S., & Manning, C. (2012). Baselines and Bigrams: Simple, Good Sentiment and Topic Classification. *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics* (fv. 90–94). Jeju, Republic of Korea: Association for Computational Linguistics.

Webb, G. (2011). Model Evaluation. Në C. Sammut, & G. Webb, *Encyclopedia of Machine Learning.* Boston, MA.: Springer.

Westgate, A., & Valova, I. (2018). A Graph Based Approach to Sentiment Lexicon Expansion. *IEA/AIE 2018: Recent Trends and Future Technology in Applied Intelligence* (fv. 530-541). Springer.

Witten, H., Frank, E., Hall, A. M., & Pal, J. C. (2017). *Data Mining, Practical Machine Learning Tools and Techniques* (Vol. 4th Edition). Elsevier Inc.

Xu, H., Liu, B., Shu, L., & Yu, P. S. (2018). Double Embeddings and CNN-based Sequence Labeling for Aspect Extraction. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (fv. 592–598). Melbourne, Australia: Association for Computational Linguistics.

Xu, J., Chen, D., Qiu, X., & Huang, X. (2016). Cached long short term memory neural networks for document level sentiment classification. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, (fv. 1660–1669).

Zeman, D., Hajič, J., Popel, M., Potthast, M., Straka, M., Ginter, F., . . . Petrov, S. (2018). CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies. *The CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies.*

Zhang, L., Wang, S., & Liu, B. (2018). Deep Learning for Sentiment Analysis : A Survey. *WIREs Data Mining Know Discovery*.

Zheng, X. (2017). Incremental Graph-based Neural Dependency Parsing. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing* (fv. 1655-1665). Copenhagen, Denmark: Association for Computational Linguistics.