REPUBLIKA E SHQIPËRISË
UNIVERSITETI POLITEKNIK I TIRANËS
FAKULTETI INXHINIERISË MATEMATIKE DHE INXHINIERISË FIZIKE
DEPARTAMENTI INXHINIERISË MATEMATIKE

DISERTACION

Për marrjen e gradës shkencore
"DOKTOR"
Paraqitur nga

MSc. Denisa SALILLARI (KAÇORRI)

MODELIMI MATEMATIK I GERMAVE NË TEKSTET SHQIP DHE ZBATIME NË SISTEME KOMPJUTERIKE

Udhëheqës Shkencor

Prof. Asoc. Dr. Luela PRIFTI

Tiranë, 2016

REPUBLIKA E SHQIPËRISË
UNIVERSITETI POLITEKNIK I TIRANËS
FAKULTETI INXHINIERISË MATEMATIKE DHE INXHINIERISË FIZIKE
DEPARTAMENTI INXHINIERISË MATEMATIKE

DISERTACION

Për marrjen e gradës shkencore
"DOKTOR"
Paraqitur nga

MSc. Denisa SALILLARI (KAÇORRI)

PROGRAMI I STUDIMIT: Statistikë dhe Kërkime Operacionale.

TEMA:

MODELIMI MATEMATIK I GERMAVE NË TEKSTET SHQIP DHE ZBATIME NË SISTEME KOMPJUTERIKE

Udhëheqës Shkencor: Prof. Asoc. Dr Luela PRIFTI

Mbrohet më datë ___/___/ 2016 para jurisë:

1. Prof. Dr. Shpetim SHEHU Kryetar, Oponent
2. Prof. Dr. Viktor KABILI Anëtar, Oponent
3. Prof. Dr. Agron TATO Anëtar
4. Prof. Dr. Shpetim LEKA Anëtar
5. Prof. Dr. Nevruz KOÇI Anëtar

Dedikuar Familjes

Falenderime!

I jam thellësisht mirënjohëse dhe e falenderoj udhëheqësen time shkencore Prof. Asoc. Dr. Luela Prifti për udhëheqjen, bashkëpunimin dhe mbështetjen, e saj shkencore e miqësore gjatë të gjithë kohës së përgatitjes të këtij disertacioni!

Falenderoj profesorin e nderuar Prof. Dr Shpetim Shehu për nxitjen, këshillat dhe sugjerimet për të nisur dhe finalizuar me sukses këtë studim.

Falenderoj kolegët e mi të departamentit të Inxhinierisë Matematike për mbështetjen dhe këshillat e tyre. Në veçanti falenderoj kolegun e departamentit të Inxhinierisë Informatike në UPT Dr. Hakik Paci për ndihmën dhe bashkëpunimin duke bërë të mundur ndërtimin e programit me të cilin u mundësua realizimi i aplikimit në këtë studim.

I detyrohem shumë familjes time, veçanërisht prindërve të mi dhe i jam mirënjohëse përjetë për ndihmën me përkushtim e dashuri në shtytjen drejt rrugës së dijes. Falenderim i veçantë shkon për bashkëshortin tim Petrit dhe për " bashkëpunëtoren" më të rëndësishme vajzën time Alsea, për dashurinë, durimin dhe qetësinë që më kanë dhuruar.

PËRMBAJTJA

LISTA E FIGURAVE III

LISTA E TABELAVE IV

PËRMBLEDHJE 1

ABSTRACT 3

KAPITULLI 1 4

ZINXHIRËT E MARKOVIT NË TEKSTET SHQIP 4

1.1 HYRJE 4
1.2 ZINXHIRËT E MARKOVIT 5
1.3 PROBABILITETET E KALIMEVE TË RENDEVE TË LARTA, EKUACIONET CHAPMAN-KOLMOGOROV 7
1.4 MODELI I ZINXHIRËVE TË MARKOVIT PËR IDENTIFIKIMIN E TEKSTEVE 7
1.5 NDËRTIMI I PROGRAMIT 11
1.5.1 Analiza e database-it 11
1.5.2 Analiza logjike e metodës 12
1.6 APLIKIMI I METODËS NË TEKSTET E HUAJA 14
1.6.1 Projekti Gutenberg 14
1.6.2 Jashtë guvës së hijeve 14
1.6.3 Shkrimet Federale 15
1.6.4 Diskutime dhe përfundime 16
1.7 APLIKIMI I METODËS NË TEKSTET E GJUHËS SHQIPE 17
1.7.1 Rezultatet në rastin e parë të studimit 18
1.7.2 Rezultatet në rastin e dytë të studimit 22
1.8 PËRFUNDIME 28

KAPITULLI 2 29

ANALIZE STATISTIKORE E TEKSTEVE NË GJUHËN SHQIPE 29

2.1 HYRJE 29
2.2 LIGJI ZIPF 29
2.3 ANALIZA E GRUPIMEVE 30
2.4 GJUHA R DHE PAKETAT PËR PËRPUNIMIN E TEKSTEVE 31
2.5 ANALIZA STATISTIKORE E TEKSTEVE SHQIP NË R 31
2.6 SHPËRNDARJA E GERMAVE DHE E FJALËVE NË TEKSTET SHQIP 40
2.7 VARIABLA SASIORE NGA TEKSTET SHQIP 45
2.9 PËRFUNDIME 48

KAPITULLI 3 49

REGRESI LOGJISTIK 49

3.1 HYRJE 49
3.2 REGRESI I THJESHTË LOGJISTIK 49
3.2.1 Përshtatja e Modelit të Regresit Logjistik 51
3.2.2 Kontrolli i Rëndësise se Koefiçentëve dhe intervalet e besimit 52
3.3 REGRESIONI LOGJISTIK I SHUMFISHTË 54
3.3.1 Përshtatja e Modelit të Regresit Logjistik të Shumëfishtë 55
3.3.2 Kontrolli i Rëndësisë së Modelit 56
3.4 REGRESI LOGJISTIK MULTINOMIAL 57
3.5 MODELET E REGRESIT LOGJISTIK PËR IDENTIFIKIMIN E AUTORIT 58
3.5.1 Regresi i shumëfishtë logjistik 58
3.5.2 Regresi logjistik multinomial 59
3.6 MODELI I REGRESIT LOGJISTIK PËR TEKSTE NË GJUHËN SHQIPE 60

3.6.1 Variablat për tekstet në gjuhën shqipe *60*
3.6.2 Modelet e regresit logjistik të shumëfishtë për tekset shqip *60*
3.7 MODELI I REGRESIT LOGJISTIK MULTINOMIAL PËR TEKSTET NË GJUHËN SHQIPE 71
3.8 PËRFUNDIME 104

LITERATURA 106

Lista e figurave

FIGURA 1: Pema e familjes së gjuhëve indoevropiane. Burimi: http://www.linguatics.com/indoeuropean_languages.htm 1

FIGURA 1. 1: Paraqitja e Programit..... 10

FIGURA 1. 2: Skema e database-it tone..... 12

FIGURA 1. 3: Bllokskema e algoritmit te programit 13

FIGURA 1. 4: Llogaritja e frekuencave të shfaqjes së një gërme pas një germe tjetër, mbi bazën e së cilës ndërtohet matrica Markoviane e rendit të parë. 18

FIGURA 1. 5: Caktimi i rendit të Zinxhirit të markovit. 19

FIGURA 1. 6: Rezultatet e koeficientëve të entropisë sipas secilit autor, për tekstin e zgjedhur..... 19

FIGURA 1. 7: Matrica e frekuencave Per tekstin Tekst14. 23

FIGURA 1. 8: Rezultatet per Zinxirin e Markovit të rendit të pare per TEKST14. 24

FIGURA 2. 1: Fjala Për si zinxhir Markovi 35

FIGURA 2. 2: Grafiku i shperndarjes së fjalëve më të shpeshta shqip 37

FIGURA 2. 3: Reja e fjalëve në korpusin shqip 37

FIGURA 2. 4: Reja e fjalëve në korpusin shqip me frekuencë të paktën 700 38

FIGURA 2. 5: Dendograma e fjaleve ne korpusin shqip..... 39

FIGURA 2. 6: Ligji Zipf per fjalet ne korpusin shqip 40

FIGURA 2. 7: Ligji Zipf per fjalet ne korpusin artikuj 41

FIGURA 2. 8: Ligji Zipf për germat në korpusin artikuj 44

FIGURA 2. 9 : Ligji Zipf për germat në korpusin Shqip..... 44

Lista e tabelave

TABELA 1. 1: Rezultatet e riklasifikimit 14

TABELA 1. 2: Përcaktimi i autorit për secilin Artikull në rastin e pare të studimit duke u bazuar në zinxhiret e Markovit deri në rendin e gjashtë 21

TABELA 1. 3: Përcaktimi i autorit për secilin Roman në rastin e pare të studimit duke u bazuar në zinxhiret e Markovit deri në rendin e gjashtë 21

TABELA 1. 4: Probabilitetet e suksesit në rastin e parë të studimit për secilin model... 22

TABELA 1. 5: Rezultatet e identifikimit të autoreve të artikujve në rastin e dyte të studimit. 25

TABELA 1. 6 Rezultatet e identifikimit të autoreve të romaneve në rastin e dyte të studimit. 26

TABELA 1. 7: Probabilitetet e identifikimit të autorit të teksteve në rastin e dyte të studimit 26

TABELA 2. 1: Shperndarja e shkronjave ne korpsin shqip..... 42

TABELA 2. 2: Shperndarja e shkronjave ne korpusin artikuj..... 43

TABELA 3. 1: Ligji probabilitar i madhesise **ε** 50

TABELA 3. 2: Rezultatet e klasifikimit te teksteve ne modelin 1 64

TABELA 3. 3: Rezultatet e klasifikimit ne modelin 3 te autorit 1 68

TABELA 3. 4: Rezultatet e klasifikimit te modelit 3 per secilin autor 68

TABELA 3. 5: Rezultatet e klasifikimit te teksteve ne modelin 4 te autorit 1 70

TABELA 3. 6:Rezultatet e klasifikimit te modelit 4 per secilin autor 70

TABELA 3. 7: Vleresimi i parametrave ne modelin e punimit Salillari, Prifti (2016) 71

TABELA 3. 8: Rezultatet e klasifikimit per modelin ne punimin Salillari, Prifti (2016) 72

TABELA 3. 9: Rezultatet e klasifikimit per modelin e paraqitur ne punimin Salillari, Prifti(2016)..... 76

TABELA 3. 10: Rezultatet e klasifikimit ne modelin MLARTSH1 80

TABELA 3. 11: Rezultatet e klasifikimit ne modelin MLARTSH2 84

TABELA 3. 12: Rezultatet e klasifikimit ne modelin MLARTSH4 91

TABELA 3. 13: Rezultatet e klasifikimit ne modelin MLARTSH5 95

TABELA 3. 14: Rezultatet e klasifikimit ne modelin MLARTSH6 100

PËRMBLEDHJE

Identifikimi i autorit ndeshet në problematika të ndryshme si: në njohjet e plagjiaturës, njohjen e zërit, mjekësinë ligjore, në kriminalistikë, inteligjencën ushtarake, etj. Problemi kryesor i identifikimit të autorit i cili është çështja kryesore në këtë disertacion, është identifikimi i autorësisë së teksteve. Çdo mendim apo ide e jona shprehet me anë të fjalës, nëpërmjet gjuhës së folur dhe të shkruar e cila është unike për individin. Gjuha e folur dhe e shkruar përbëhet nga germat, leksiku dhe sintaksa. Për të njohur sa më shumë karakteristika gjuhësore të një individi duhet të bëjmë modelime matematike të tyre. Gjuha shqipe është një degëzim i veçante në familjen e gjuhëve Indo-Europiane, si e tillë kërkon një trajtim matematikor të veçantë në studimin e metodave matematikore të përdorura në gjuhë të tjera për identifikimin e autorësisë së teksteve.

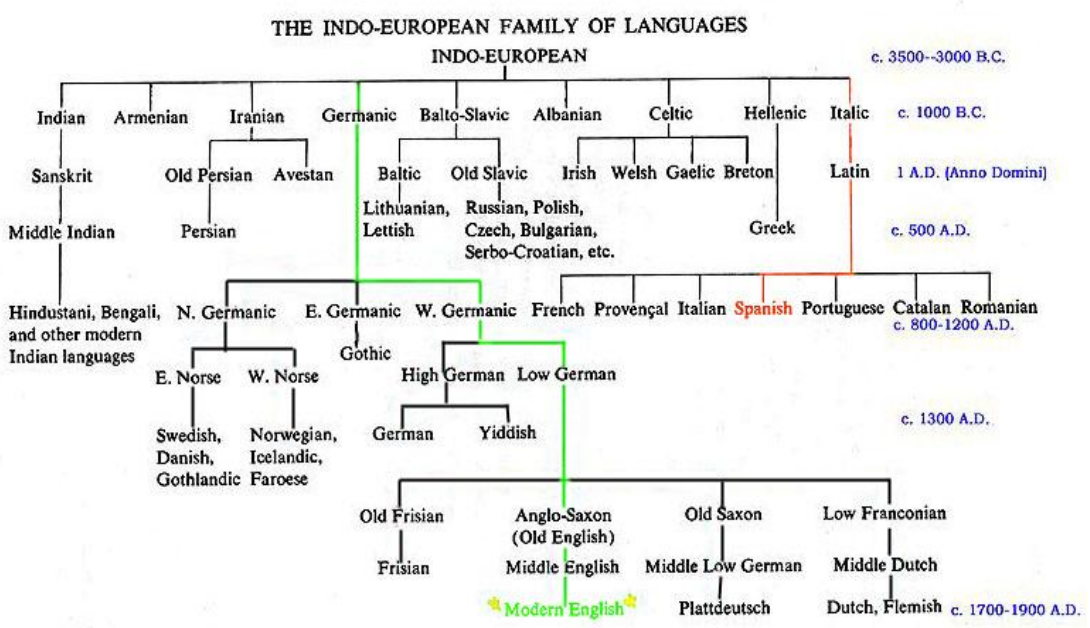

FIGURA 1: PEMA E FAMILJES SË GJUHËVE INDOEVROPIANE. BURIMI:
http://www.linguatics.com/indoeuropean_languages.htm

Qëllimi i këtij studimi është identifikimi i autorit në tekstet shqip me anë të modelimeve probabilitare dhe metodave të klasifikimit. Për herë të parë në tekstet shqip ndërtohen modelet e zinxhirëve të Markovit deri në rendin e gjashtë me një program të ndërtuar në gjuhën VB, si dhe me besueshmëri 95% ndërtohen modele të regresit logjistik në gjuhën R.

Studimi është i organizuar në tre kapituj. Në secilin kapitull përshkruhet teorikisht metodologjia e ndjekur, përmirësimet e bëra në metoda si dhe aplikimet e realizuara. Si pikë e fundit e çdo kapitulli paraqiten përfundimet e arritura.

Në kapitullin e parë të këtij studimi trajtohen modelime të teksteve në gjuhën shqipe, të bazuara në zinxhirët e Markovit, duke ndërtuar matrica Markoviane të germave deri në rendin e gjashtë. Modelet janë ndërtuar në dy raste studimi: në rastin e parë supozohet si hapësirë gjendjesh bashkësia me germat e thjeshta të alfabetit shqip dhe karakterin hapësirë, në rastin e dytë supozohet si hapësirë gjendjesh, alfabeti shqip bashkë me karakterin hapësirë. Studimi është bërë në një korpus prej 107 tekstesh, të cilët janë romane dhe artikuj të shtypit të përditshëm shqip, të 13 autorëve të ndryshëm. Modelet janë realizuar nëpërmjet një programi të ndërtuar në gjuhën VB, bazuar në zinxhirët e Markovit, sipas metodës së paraqitur nga Khmelev et al.(2001).

Në kapitullin e dytë është bërë një analizë statistikore e fjalëve dhe germave në gjuhën shqipe, nga e cila janë nxjerrë disa rezultate të rëndësishme për shpërndarjen e tyre. Janë bërë vlerësime pikësore dhe intervalore për variablat sasiore në studim: numri i germave, numri i fjalëve, numri i fjalive, numri i shenjave të pikësimit etj, të marre nga 107 tekste shqip. Në këtë kapitull kontribut është edhe përpunimi, analiza statistikore e elementeve gjuhësor dhe sintaksor të teksteve shqip me disa paketa të programit R.

Në kapitullin e tretë janë trajtuar teorikisht modelet e regresit logjistik të thjeshtë, regresit logjistik të shumëfishtë dhe regresit logjistik multinomial si edhe trajtimi i këtyre modeleve si metodë klasifikimi me qëllim identifikimin e autorit të teksteve. U ndërtuan modele të regresit logjistik të shumëfishtë dhe modele të regresit logjistik multinomial për 100 tekste shqip. Në pjesën e fundit të këtij kapitulli diskutohen modelet e ndërtuara në R, duke përcaktuar parametrat më të mirë që modelojnë tekstet shqip, si dhe janë renditur avantazhet e disavantazhet e metodave.

Abstract

The authorship's identification has been for a long time the focus of many researchers. In recent years, they gave attention to authorship analysis for plagiarism detection, voice recognition, forensic cases and authorship attribution of different texts like emails, electronic messages, articles, books etc. The main problem of identifying the author who is also the main issue in this thesis is to identify the authorship of texts. Different statistical methods are applied to identify the authorship of a text for different languages. To recognize many linguistic features of an individual, is needed the development of mathematical models of letters, vocabulary and syntax of a language. In this study the authorship attribution for Albanian texts is considered as a classification problem. The Albanian language belongs to the Indo-European family of languages and forms an independent branch of its own figure 1. The Albanian language is different from the other languages due to its difficult syntactic structure which motivate us to the application of some statistical methods in authorship identification for Albanian texts. In the first chapter of this thesis are presented different Markov chains models defined on letters for Albanian language which are successfully applied on 107 texts in Albanian language, estimating the texts entropy via Markov chains of high order. As conclusion Markov chain of second order results as the best Markov chain model for Albanian texts. In the second chapter using R is realized text mining of Albanian texts, a statistical analysis of data taken from different Albanian texts, and are defined the distribution of letters and word in Albanian texts. In the third chapter the authorship identification is considered as a classification problem through logistic regression and multinomial logistic regression models. In this chapter using R is improved the logistic regression model increasing the number of texts and number of independent variables. The application is realized with data from our Albanian text considered in previous chapters, considering as the independent variables in the models, number of letters, number of words, number of vowels, number of consonants, number of punctuations and number of sentences etc. As conclusions are discussed the best fitted model, the most significant variables and compared these models to each others.

Keywords: Authorship attribution, Markov Chain, Logistic Regression,

Kapitulli 1

ZINXHIRËT E MARKOVIT NË TEKSTET SHQIP

1.1 Hyrje

Zinxhirët e Markovit përdoren si një model për vargjet e elementeve të një teksti gjuhësor. Ky model aplikohet për njohjen e teksteve. Një element i tekstit mund të jetë një germë ose një klasë gramatikore e një fjale. Frekuenca e përdorimit të germave dy e nga dy të një klase gramatikore janë karakteristika të qëndrueshme të një autori dhe mund të përdoren në njohjen e autorësisë. Në këtë kapitull do të prezantohet një teknike për njohjen e autorësisë, bazuar në një zinxhir të thjeshtë të Markovit për germat. Khmelev(2000) dhe Khmelev, Tweedie (2001) e kanë paraqitur dhe aplikuar këtë teknike në një numër të madh tekstesh. Pavarësisht nga numri i madh i metodave Milov(1994) dhe Holmes (1998), asnjëra prej tyre nuk ishte aplikuar në një numër të madh tekstesh. Metoda të tilla pa ndihmën e softeve, e bëjnë aplikimin e tyre në një numër të madh tekstesh shpesh të pamundur. Përgjithësimi i këtyre metodave është ende i rëndësisë së veçantë.

Një metodë e cila është testuar në një numër të madh tekstesh është propozuar nga Fomenko.V, Fomenko.T (1996). Ata shqyrtuan përpjesën e fjalëve kyçe (funksion) që përdoreshin nga një autor, dhe konstatuan për një numër të madh shkrimtarësh rusë, që kjo ishte e pandryshueshme për secilin autor.

Në këtë kapitull përshkruhet teoria e zinxhirëve të Markovit, metoda e paraqitur nga Khmelev (2000) si dhe aplikimet e saj në gjuhë të huaja. Metoda e paraqitur nga Khmelev (2000) dhe Khmelev et al.(2001) përshtatet për modelime të teksteve në gjuhën shqipe, duke ndërtuar matrica Markoviane të germave deri në rendin e gjashtë. Modelet janë ndërtuar në dy raste studimi: në rastin e parë supozohet si hapësirë gjendjesh bashkësia me germat e thjeshta të alfabetit shqip dhe karakterin hapësirë, në rastin e dytë supozohet si hapësirë gjendjesh, alfabeti shqip bashkë me karakterin hapësirë. Studimi është bërë në një korpus prej 107 tekstesh, të cilët janë romane dhe artikuj të shtypit të përditshëm shqip, të 13 autorëve të ndryshëm. Modelet janë realizuar nëpërmjet një programi të ndërtuar në gjuhën VB, bazuar në zinxhirët e Markovit, sipas metodës së paraqitur nga Khmelev et al.(2001) dhe përshtatur për vlerësimin e zinxhirëve të Markovit deri në rendin e gjashtë.

1.2 Zinxhirët e Markovit

Një proces rasti është modeli matematik i një procesi empirik i cili bazohet në ligjet probabilitare Ross.Sh (2007). Proceset e rastit parashikojnë modele të dobishme për studime në fusha të ndryshme si fizikë, telekomunikacion, analize e serive kohore, rritjen e popullsisë, shkenca menaxhimi etj.

Përkufizim 1.2.1. Le të jetë (Ω, A, P) një hapësirë probabilitare e dhënë dhe X një ndryshore rasti nga kjo hapësirë probabilitare. Një bashkësi (familje) ndryshoresh rasti $\{X_t, t \in T\}$ me vlera në $\mathbb{R}$ e indeksuar nga parametri "kohë" t$\in$ T quhet proces stokastik ose proces rasti.

Pra një proces rasti është funksion me dy argument $\{X_{t,\omega}, t \in T \; dhe \omega \in \Omega\}$.

Për çdo $\omega \in \Omega$ të fiksuar, $X_t(\omega)$ është një funksion i t-së. Ai quhet një realizim i procesit stokastik ose trajektore e procesit.

Për çdo $t \in T$ të fiksuar, $X_t(\omega)$ është një funksion i ω-së. Pra $X_t(\omega)$ është një ndryshore rasti.

Për $\omega \in \Omega \; dhe \; t \in T$ të fiksuar, $X_t(\omega)$ është një numër.

Për procesin e rastit nuk do të përdorim $X(t, \omega)$ por do të shkruajmë vetëm $\{X(t), t \in T\}$ por nuk do të harrojmë se është funksion i ω-ës. Keshtu qe procesi do të përshkruhet nga dy bashkësi, T hapësira parametrike dhe E hapësira e gjendjeve.

Në qoftë se T është një bashkësi diskrete, atëherë $\{X_t, t \in T\}$ quhet proces stokastik me kohë (hapësirë parametrike) diskrete.

Nëse T është një bashkësi e vazhduar (e panumërueshme), $\{X_t, t \in T\}$ do të quhet një proces stokastik me kohë (hapësirë parametrike) të vazhdueshme.

Vlerat e $X(t)$ quhen gjendje, të gjitha vlerat e ndryshme të X(t) formojnë hapësirën e gjendjeve të rastit e cila shënohet me E.

Në qoftë se hapësira e gjendjeve E, është diskrete atëherë procesi njihet si proces me hapësirë gjendjesh diskrete. Në qoftë se hapësira e gjendjeve E, është e vazhduar procesi quhet me hapësirë gjendjesh të vazhduar.

Përkufizim 1.2.2: Procesi i rastit $\{X(t), t \in T\}$ quhet proces Markovi në qofte se

$$P\left(X_{t_{n+1}} \leq x_{n+1} / \; X_{t_1} = x_1, X_{t_2} = x_2, \dots, X_{t_n} = x_n\right) = P\left(X_{t_{n+1}} \leq x_{n+1} / X_{t_n} = x_n\right)$$

për $0 < t_1 < t_2 < \cdots < t_n < t_{n+1}$

Kjo quhet vetia Markoviane ose vetia e humbjes së kujtesës.

Përkufizim 1.2.3: Procesi i Markovit me hapësirë gjendjesh diskrete quhet zinxhir Markovi.

Pra zinxhir Markovi quhen vargjet e ndryshoreve të rastit në të cilët gjendja e sistemit në një moment të dhënë n+1 nuk varet nga e kaluara e procesit por vetëm nga gjendja e sistemit në momentin n.

Një trajte tjetër e përkufizimit 1.2.3 jepet si më poshtë:

Përkufizim 1.2.3': Një varg $(X_t)_{t\geq 0}$ ndryshoresh rasti pozitive me vlera në një hapësirë të numërueshme E quhet zinxhir markovian po qe se $\forall t \geq 0, \forall\, a_0, a_1, \dots, a_{t+1} \in E$

$$P(X_{t+1} = a_{t+1} / X_t = a_t, \dots, X_0 = a_0) = P(X_{t+1} = a_{t+1} / X_t = a_t)$$

me kusht që $P(X_t = a_t, \dots, X_0 = a_0) \neq 0$.

Përkufizim 1.2.4: Madhësia $p_{ij} = P(X_{n+1} = j / X_n = i)$ quhet probabilitet kalimi me një hap nga gjendja i në gjendjen j në momentin n.

Përkufizim 1.2.5: Një zinxhir Markovi quhet homogjen po qe se probabilitetet e kalimit nuk varen nga koha n.

Në një zinxhir Markovi homogjen madhësitë p_{ij} janë numra nga 0 në 1 dhe i vendos në një matricë. Kjo matricë i ka të gjithë elementët nga 0 në 1 dhe shuma e elementëve të çdo rreshti është 1. Kjo matricë që gëzon vetitë e mësipërme quhet matricë Markoviane ose matricë kalimi.

Përkufizim 1.2.6: Le të jetë $A = \{a_1, a_2, \dots, a_n\}$ bashkësia e të gjitha vlerave të mundshme që merr një madhësi e rastit X dhe $P = \{p_1, p_2, \dots, p_n\}$ shpërndarja probabilitare e kësaj madhësie rasti me $P(x = a_i) = p_i$. Bashkësia A quhet alfabet.

Informacioni Shannon (Cloude Shannon 1948) që përmban një rezultat (një ngjarje) x i madhësisë së rastit X është

$$h(x) = \log_2 \frac{1}{P(x)}$$

njësia matëse e informacionit është bit.

Në vend të logaritmit me bazë 2 në teorinë e informacionit merren edhe logaritmet me bazë 3, e=2.718, në këto raste njësitë matëse të informacionit do të jenë përkatësisht trits (për "trinary") dhe nats(për "naturals")

Përkufizim 1.2.7: Entropi e një madhësie rasti quhet mesatarja e informacionit Shannon që përmban madhësia e rastit X

$$H_P(X) = -\sum_{a_i \in A} P(X = a_i) ln P(X = a_i) = \sum_{i=1}^{n} p_i \ln (1/p_i)$$

Entropi për një madhësi X do të thotë informacioni i panjohur (pasiguri) i madhësisë X.

1.3 Probabilitetet e kalimeve të rendeve të larta, ekuacionet Chapman-Kolmogorov

Shpërndarja probabilitare e një zinxhiri Markovi përcaktohet me ndihmën e veprimeve të matricave. Shënojmë $P = [p_{ij}]$ matricën e probabilitetit të kalimeve të një zinxhiri Markovi. $P^2 = PP$ me element në rreshtin *i* dhe shtyllën *j*: $p_{ij}^2 = \sum_k p_{ik} p_{kj}$.

Për hapësira gjendjesh të pafundme seria e mësipërme konvergjon pra:

$$\sum_k p_{ik} p_{kj} < \sum_k p_{ik} = 1$$

Në të njëjtën mënyrë tregohet se $P^3 = PP^2$ me element në rreshtin *i* dhe shtyllën *j*: $p_{ij}^3 = \sum_k p_{ik} p_{kj}^2$. Në përgjithësi kemi $P^{n+1} = PP^n$ me element në rreshtin *i* dhe shtyllën *j*: $p_{ij}^{n+1} = \sum_k p_{ik} p_{kj}^n$ ku $p_{ij}^n = P(X_n = j / X_0 = i)$. Pra probabilitetet e kalimit me n hapa nga një gjendje i në një gjendje j llogariten nëpërmjet llogaritjes së fuqive të matricës P.

Identiteti i matricave $P^{m+n} = P^m P^n$ për m>0 dhe n>0 njihet si ekuacioni Chapman-Kolmogorov. Ekuacioni Chapman-Kolmogorov shpreh faktin që kalimi nga gjendja i në gjendjen j me m+n hapa arrihet duke kaluar me m hapa nga gjendja i drejt një gjendje të ndërmjetme k me probabilitet p_{ik}^m dhe me pas duke kaluar nga gjendja k drejt gjendjes j me n hapa me probabilitet p_{kj}^n, pra ngjarjet "shko nga i në k me m hapa" dhe "shko nga k në j me n hapa" janë ngjarje të pavarura.

1.4 Modeli i zinxhirëve të Markovit për identifikimin e teksteve

Në këtë paragraf do të prezantohet një teknikë për njohjen e autorësisë, të bazuar në një zinxhir të thjeshtë të Markovit për germat Khmelev et al. (2001). Në disa shembuj të thjeshtë janë ilustruar disa metoda që sugjerohen nga Khmelev et al. (2001) për njohjen e autorësisë së teksteve në gjuhë të huaj. Metoda jep rezultate të shkëlqyera kur aplikohet në të paktën 380 tekste. Për gjetjen e një modeli probabilitar për gjuhën e përdorur në tekste, duhet të konsiderojmë një model ku germat dhe hapësirat gjenerohen në mënyrë të pavarur, duke pasur parasysh probabilitetet e përsëritjes së germave. Sigurisht germat nuk vendosen në mënyrë të rastësishme dhe janë të varura nga germat që përsëriten përpara

tyre. Modele të tilla të thjeshta kërkojnë që secila germë të varet vetëm nga germa përpara saj. Tregohet që këtë model mund ta përdorim për të përcaktuar autorësinë në një gamë të gjerë shembujsh. Zinxhirët e Markovit përdoren si një model për vargjet e elementeve të një teksti gjuhësor, ky model aplikohet për njohjen e teksteve. Një element i tekstit mund të jetë një germë ose një klasë gramatikore e një fjale. Metoda tregon se frekuenca e përdorimit të germave dy e nga dy të një klase gramatikore janë karakteristika të qëndrueshme të një autori dhe mund të përdoren në njohjen e autorësisë. Pra një varg germash konsiderohet si një zinxhir Markovi ku secila germe varet vetëm nga germa përpara saj.

Një shembull i thjeshtë i zinxhirit të Markovit mund të ilustrohet në një sekuencë të një numri të kufizuar germash Salillari.D et al. (2012) , psh: të një shprehje fjalëshpejtë

"Kupa me kapak, kupa pa kapak."

Në këtë fjali kemi gjashtë germa të ndryshme A, E, K, M, P, U. Në qoftë se numërojmë se sa herë germa A shfaqet në këtë sekuencë të ADN vërejmë që germa A shfaqet 7 herë, germa A pasohet nga germa K në 3 raste, ose 42.86% të rasteve, ndërsa pasohet nga germa M vetëm 1 herë (14.28%) , Në këtë mënyrë në mund të ndërtojmë një matricë të plotë kalimi si më poshtë:

$$F = \begin{matrix} A \\ E \\ K \\ M \\ P \\ U \end{matrix} \begin{bmatrix} 0 & 0 & 3 & 1 & 3 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 2 & 0 & 1 & 0 & 0 & 2 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 5 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 2 & 0 \end{bmatrix}$$

Matrica F kthehet në një matricë probabilitetesh duke pjesëtuar në çdo rresht secilin numër me shumën e rreshtit përkatës.

$$P = \begin{matrix} A \\ E \\ K \\ M \\ P \\ U \end{matrix} \begin{bmatrix} 0.0000 & 0.0000 & 0.4286 & 0.1428 & 0.4286 & 0.0000 \\ 0.0000 & 0.0000 & 1 & 0.0000 & 0.0000 & 0.0000 \\ 0.4 & 0.0000 & 0.2 & 0.0000 & 0.0000 & 0.4 \\ 0.0000 & 1 & 0.000 & 0.0000 & 0.0000 & 0.0000 \\ 1 & 0.0000 & 0.000 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 0.0000 & 0.000 & 0.0000 & 1 & 0.0000 \end{bmatrix}$$

Matrica P është një matricë Markoviane.

Në gjuhën shqipe alfabeti ka 36 shkronja, duke shtuar edhe karakterin hapësirë, për tekstet në gjuhën shqipe do të ndërtohet një matricë Markoviane me përmasa 37x37.

Një matricë kalimi e krahasueshme me atë më sipër megjithëse më e madhe llogaritet për secilin tekst. Pra për çdo tekst llogaritet matrica e kalimit me frekuencat e kalimit nga një germë në një tjetër. Matrica e kalimit për një autor përftohet duke llogaritur shumat e matricave të kalimit të teksteve të tij përkatëse.

Në këtë mënyrë llogaritet me përafërsi probabiliteti i kalimit nga një germë në një tjetër për secilin autor. Zgjedhim një tekst anonim të këtyre autoreve. Për të përcaktuar autorësinë e këtij teksti duke supozuar se ai është shkruar nga një prej autorëve të njohur konsiderohet probabiliteti që teksti të jetë i gjeneruar nga secila prej matricave të kalimit. Përcaktimi i autorit të vërtetë do të behet duke përdorur parimin e përgjasisë maksimale. Pra secilit autor do ti vendoset në korrespondencë një probabilitet. Këto probabilitete më pas renditen, autori me probabilitetin më të madh renditet me numrin 0, tjetri me numrin 1 edhe kështu me radhë, dhe autori me probabilitetin më të madh pra dhe me renditjen më të vogël mendohet të jetë ai i cili ka shkruar tekstin. Për këto probabilitete llogariten logaritmet natyrore të tyre duke u ndryshuar shenjën dhe duke pjesëtuar me gjatësinë e tekstit anonim. në këtë mënyre kemi llogaritur koeficientet empirike të entropisë. Autori i vërtetë i tekstit anonim mendohet të jete ai qe ka entropinë me të vogël.

Kjo metodë është paraqitur nga Khmelev (2000) të cilën e ka aplikuar këtë teknikë në autorët Rusë. Ai tregon se kjo teknikë jep rezultate shumë më të mira se analiza e artikujve (germave, shkrimeve) individuale. Më poshtë paraqitet përshkrimi matematik të metodës.

Le të jenë A autorë tekstesh ku secili ka N_a tekste, ku $a = 0, 1\ldots,A\text{-}1$, shënojmë Q_{ij}^{an} numrin e kalimeve nga germa i në germën j, për tekstin n ($n=0,\ldots,N_{a-1}$) nga shkrimtari a ($a = 0,\ldots A\text{-}1$). Për të gjetur autorin e parashikuar në tekstin $\hat{\boldsymbol{n}}$ të autorit $\hat{\boldsymbol{a}}$ përdorim informacionin e autorësisë të gjithë teksteve të të gjithë autorëve të tjerë duke përjashtuar autorin $\hat{\boldsymbol{a}}$ atëherë kemi:

$$Q_{ij}^{k} = \sum_{n=0}^{Na-1} Q_{ij}^{kn} \quad ; \qquad Q_{i}^{k} = \sum_{j} Q_{ij}^{k} \qquad \text{për autorët} \quad k \neq \hat{a}$$

dhe për autorin $\hat{\boldsymbol{a}}$ duke përjashtuar tekstin $\hat{\boldsymbol{n}}$ nga bashkësia e trajtuar kemi:

$$Q_{ij}^{\hat{a}} = \sum_{n \neq \hat{n}} Q_{ij}^{\hat{a}n} \quad ; \qquad Q_{i}^{\hat{a}} = \sum_{j} Q_{ij}^{\hat{a}}$$

Atëherë do të kemi

$$\Lambda_k(\hat{a},\hat{n}) = -\sum_{i:Q_i^k>0} \; \sum_{j:Q_{i,j}^k>0} Q_{ij}^{\hat{a}\hat{n}} \ln\left(\frac{Q_{ij}^k}{Q_i^k}\right)$$

dhe

$$\Lambda_{\hat{a}}(\hat{a},\hat{n}) = -\sum_{i:Q_i^{\hat{a}}>0} \sum_{j:Q_{i,j}^{\hat{a}}>0} Q_{ij}^{\hat{a}\hat{n}} ln\left(\frac{Q_{ij}^{\hat{a}}}{Q_i^{\hat{a}}}\right)$$

Në qoftë se $Q_{ij}^{k}=0$ dhe $Q_i^k=0$ atëherë nuk e llogarisim këtë pjesë të shumës duke e marrë të barabartë me zero. Gjithashtu përcaktojmë rangjet $R_k(\hat{a},\hat{n})$ si rangun e $\Lambda_k(\hat{a},\hat{n})$ në $\{\Lambda_k(\hat{a},\hat{n}), k=0,....,A-1\}$ ku $R_k(\hat{a},\hat{n})\in\{0,...,A-1\}$ dhe $\Lambda_k(\hat{a},\hat{n})$ më e vogël ka rangun më të vogël. Në qoftë se teksti i përcaktohet autorit të saktë atëherë $R_{\hat{a}}(\hat{a},\hat{n})=0$. Duke u bazuar tek algoritmi i mësipërm tek Paci et al (2011) është përshkruar ndërtimi i një programi në gjuhën VB i cili është përdorur për vlerësimin e koeficienteve të entropisë për tekstet në gjuhën Shqipe në Salillari.D et al (2012).

Gjatesia mesatare e fjalës në gjuhën shqipe është vlerësuar 6-7 karaktere nga Çiço,Cepa(2003). Për të vlerësuar probabilitetin e shfaqjes së një fjale pas një fjale të caktuar, duke ditur një vlerësim për gjatësinë mesatare të fjalës në gjuhën shqipe duhet të llogaritim probabilitetet e kalimit me 5-6 hapa nga një në një germë tjetër. Kështu që metoda e përshkruar më sipër zbatohet edhe për matricat e kalimeve të teksteve nga rendi i parë deri në rendin e gjashtë. Matricat e kalimeve të rendeve të larta për secilin tekst përftohen me ndihmën e barazimit Chapman Kolmogorov.

Programi u përmirësua duke llogaritur koeficientet e entropisë bazuar në matrica Markoviane nga rendi i parë në rendin e gjashtë. Paraqitja e programit jepet në figurën 1.1.

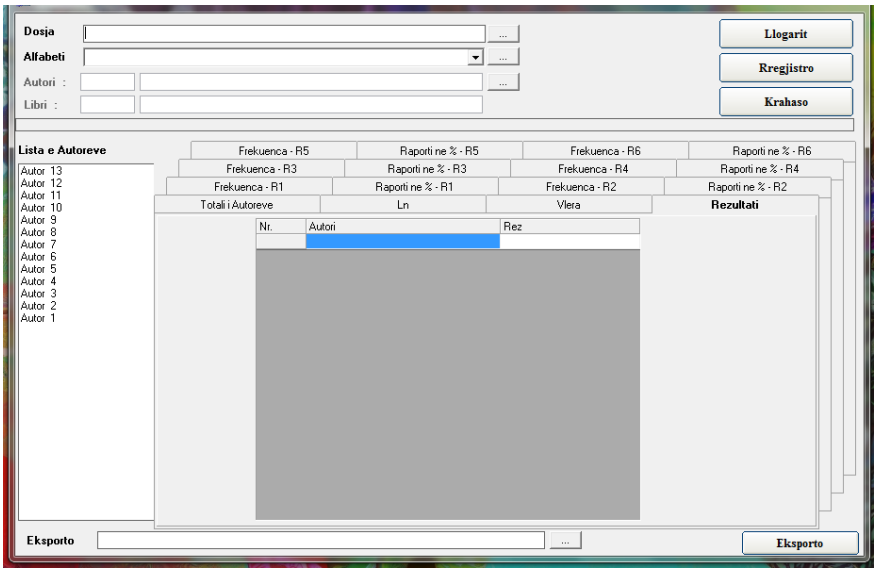

FIGURA 1. 1: PARAQITJA E PROGRAMIT.

1.5 Ndërtimi i programit

Nisur nga sasia e madhe e veprimeve për të analizuar të dhënat si dhe nga mundësitë që na ofron teknologjia e sotme është më e thjeshtë që ndërtojmë një program i cili do të na realizojë qëllimin këtij studimi.

Më poshtë po paraqesim kërkesat që duhet të realizojë ky program si dhe logjikën se si duhet të ndërtohet në mënyrë të tillë që e gjithë metoda matematikore të jetë e (zbatueshme) implementuar pa gabime :

Ndërtimi i një database i cili do të mundësojë që të dhënat e proceduara njëherë të kemi mundësinë ti ri përdorim sa herë që të jetë e nevojshme pa pasur nevojë të një riproçesimi të ri.

Në këtë database informacioni me të cilin do të popullohet duhet të jetë i mirë strukturuar në mënyrë të tillë që çdo tekst i cili do të proçesohet të mbahet i klasifikuar për autorin si dhe për alfabetin për të cilin në kemi bërë këtë analizë.

- Ndërtimi i matricës se kalimit bazuar në Alfabetin e zgjedhur.
- Ndërtimi i matricave fuqi të matricës së kalimit.
- Numër i pa limituar autoresh dhe alfabetesh.
- Mundësi eksportimi të të dhënave.
- Mundësi përdorimi i të dhënave që kemi në database për llogaritjet e metodës.
- Implementimi i metodës matematikore gjatë procedurës se krahasimeve.

1.5.1 Analiza e database-it

Database për të cilin në kemi nevoje duhet të organizohet në tabelat e mëposhtme :

Tabela e Autorëve, në të cilën do të mbahen disa informacione mbi autorët të cilët do të përfshihen në studimin tonë.

Tabela e Alfabeteve në të cilën po e emërtojmë CELES.

Tabela e Librave e cila do të popullohet me të dhënat e librave të cilët janë marrë në studim.

Tabela e Data, kjo tabele do të mbajë të gjithë matricat e kalimeve të të gjitha rendeve nga i pari deri në të gjashtin për çdo libër dhe për çdo Alfabet për të cilën është analizuar

Figura në vijim paraqet skemën e database-it tonë.

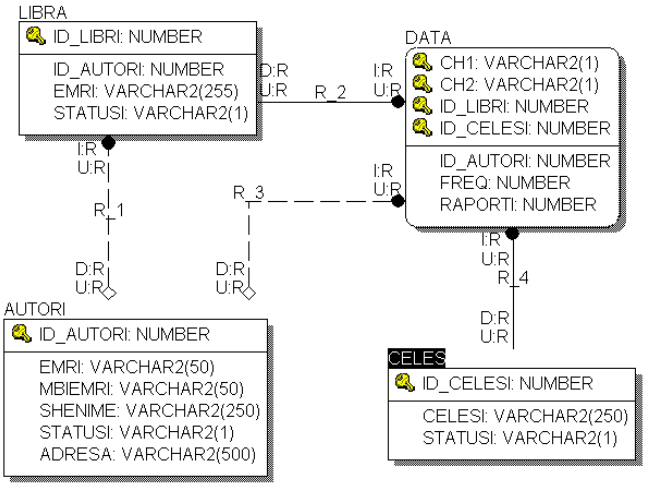

FIGURA 1. 2: SKEMA E DATABASE-IT TONE.

1.5.2 Analiza logjike e metodës

Në fillim programi duhet të llogarisë matricat e kalimit për një alfabet dhe tekstin në fjalë. Për çdo alfabet dimensionet e matricës së kalimit do të jenë sa numri i karaktereve që përmban ky alfabet. Popullimi i matricës bëhet në mënyrë shumë të thjeshtë, lexohet çdo karakter i tekstit dhe nëse ky karakter është pjesë e alfabetit tonë atëherë qeliza e matricës së kalimit me indekse karakteri pasardhës dhe karakteri i sapo lexuar rritet me 1. Kjo procedurë përsëritet deri në momentin që në i kemi lexuar të gjitha karakteret që gjenden në tekst. Theksojmë që në fillim të proçesimit të tekstit, vlerat e të gjithë qelizave të matricës së kalimit janë 0.

Pasi kemi përfunduar me leximin e tekstit, pra në momentin që në e kemi të populluar matricën e kalimeve atëherë llogaritim probabilitetin e çdo qelize të çdo rreshti, duke pjesëtuar vlerën e qelizës se matricës të sapo formuar me totalin e rreshti në të cilin bën pjesë kjo qelizë.

Nëse këto të dhëna të cilat merren nga dy procedurat e sipër përmendura janë të identifikuara se cilit autor i përket, atëherë këto të dhëna në mund të përdoren për popullimin e database-it të ndërtuar më sipër. Mënyra me të cilën do të popullohet database-i është pjesë teknike, në rastin tonë do përdorim teknologjinë ADO.

Pasi behet popullim në mënyrë të mjaftueshme atëherë bëhet e mundur analiza mbi tekste të ndryshëm bazuar në logjikën e metodës. Diagrama e mëposhtme paraqet algoritmin e zgjidhjes logjike të metodës sonë.

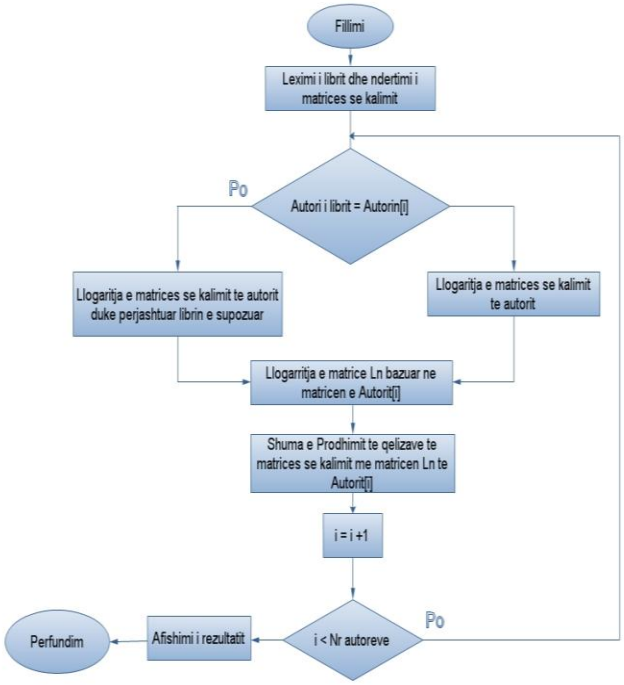

FIGURA 1. 3: BLLOKSKEMA E ALGORITMIT TE PROGRAMIT

Në arsyetimin e mësipërm nuk mund të përfshihen shkronjat që përbëhen nga 2 karaktere të cilat janë pjesë e alfabetit të gjuhës shqipe, që në këtë rast janë pjesë përbërëse e informacionit që përmban një tekst bazuar mbi konceptet e teorisë së informacionit. Pra megjithëse shkronjat e përbëra të alfabetit të gjuhës Shqipe përfitohen si kombinim i 2 kodeve ASCII, nuk mund të merret informacioni që përmban libri në mënyrë dixhitale sepse gjatë llogaritjes të matricës së kalimit do të kishim një rritje artificiale të qelizave që kanë indekse 2 karakteret e germës së përbërë.

Për të eliminuar këtë problem bëhet një zgjidhje e tillë, zëvendësohet çdo shkronjë e përbërë e alfabetit të gjuhës Shqipe me një kod ASCII i cili nuk përdoret në Alfabetin e gjuhës Shqipe dhe ky karakter vendos në alfabetin tonë të kërkimit. Me poshtë po e ilustrojmë me një shembull,

Psh : Gjuha Shqipe = 2uha 6qipe

Ku Gj = 2 dhe Sh = 6 , në alfabetin tonë duhet të shtohen edhe dy simbolet e përdorura mësipër.

1.6 Aplikimi i metodës në tekstet e huaja

1.6.1 Projekti Gutenberg

Nga arkiva e Projektit Gutenberg janë marrë tekste, në mënyrë që të shqyrtohet një numër i konsiderueshëm shkrimtarësh që shkruajnë në gjuhën angleze. Janë marrë 387 tekste nga 45 autorë të ndryshëm të cilët kanë arkivuar më shumë se një tekst. Për të bërë një test fillestar përzgjidhet në mënyrë rastësore një tekst nga secili autor. Rezultatet nga krahasimi i këtyre teksteve janë dhënë në dy kolonat e para të tabelës 1. 33 tekste janë klasifikuar siç duhet, ndërkohë që pesë tekste të tjerë kishin shkallën një, autori i duhur ishte i zgjedhuri i dytë. Mesatarja e vlerave të klasifikuara është $E(R_k) = 1.681$. Duke iu referuar numrit të teksteve 45 nga të cilat 73.3% e tyre janë të klasifikuara në mënyrë të rregullt ato shfaqin një koeficient gabimi prej 0.687%.

Një tekst		Të gjitha tekstet	
Renditja	Numri	Renditja	Numri
0	33	0	288
1	5	1	26
2	0	2	10
3	1	3	5
4	2	4	9
> 4	5	> 4	49

TABELA 1. 1: REZULTATET E RIKLASIFIKIMIT

Është bërë një riklasifikim i plotë, ku secili tekst është përjashtuar nga analiza, dhe autori i tekstit është gjykuar nga të dhëna të tjera. Nga 387 tekste dhe 45 autorë të mundshëm, janë riklasifikuar në mënyrë të rregullt 288 tekste. Rezultatet e plota janë paraqitur në pjesën e dytë të tabelës 1. Riklasifikimi mesatar i teksteve është $E(R_k) = 2.100$.

1.6.2 Jashtë guvës së hijeve

Në një studim, Baayen et al. (1996) për të përmirësuar diskriminimin e lejuar ndaj autorëve duke përdorur të dhëna sintaksore, ekzaminoi dhjetë shembuj tekstesh për secilin nga dy autorët e njohur Innes dhe Allingham. Nga njëzet shembuj origjina e katërmbëdhjetë prej tyre u njoh në ekspëriment, 6 të mbeturat duhet t'i caktohen njërit nga të dy autorët. Baayen et al. (1996) përdor analiza me komponentë themelor të fjalëve

që përsëriten vazhdimisht, matës të pasurisë leksikore dhe konstruksionet e rralla për të identifikuar autorët e gjashtë teksteve të mbetura. Ndërsa Baayen et al. (2002) aplikojnë metodat e tyre për sa i përket anës leksikore dhe sintaksore të teksteve, Khmeleve et al (2001) do të marrin në konsideratë vetëm aspektin leksikor meqenëse matrica e kalimit e rregullave sintaksore do të ishte shumë e vështirë për tu llogaritur. Dy grupet e teksteve të cilësuara më sipër formojnë bashkësinë përgatitore (trajnimit) ndërsa grupi i papercaktuar formon bashkësinë e kontrollit (test).

Krahasimi jep rezultate shumë të mira; të gjitha tekstet e autores A (Innes) u klasifikuan ashtu sikurse ishin të Innes, dhe tekstet e autores B (Allingham) u klasifikuan sikurse ishin të Allingham. Shpërndarja e këtyre gjashtë teksteve për tu klasifikuar dhe përcaktimi i tyre në mënyrë korrekte i treguar në kllapa, është përkatësisht A (A), B (B), A (B), A (A), B (B) dhe A (A). Pesë nga këto shembuj u përcaktuan në mënyrë të saktë, ky është pothuajse i ngjashëm me rezultatet e raportuara nga Baayen et al. (1996), ku si njësi matëse përdoret pasuria leksikore (katër nga gjashtë u klasifikuan në mënyrë korrekte) dhe nga përsëritja e fjalëve (pesë nga gjashtë u klasifikuan në mënyrë korrekte). Llogaritja e probabiliteteve e realizuar nga matrica e kalimit tregon që diferenca midis atributeve ka një mesatare prej 0.00060 dhe një devijim standard prej 0.0028. Teksti i tretë, i cili u klasifikua gabim, na çon në një diferencë 0.00038. Kjo teknikë përdoret ashtu si dhe disa nga metodat e aplikuara tek të dhënat leksikore nga Baayen.

1.6.3 Shkrimet Federale

Artikujt federalë janë shkruar në 1787 dhe 1788 për të bindur qytetarët e shtetit të New York të adoptonin lindjen e kushtetutës së shteteve të bashkuara. Janë 85 tekste nga të cilat 52 janë të shkruara plotësisht nga Hamilton dhe 14 tërësisht të shkruar nga Madison dhe 12 artikuj të diskutuar midis këtyre dy autorëve. Për më tepër tre tekste janë shkruar së bashku nga Hamilton dhe Madison. Pjesën e mbetur të teksteve që janë shkruar nga Jay në nuk do ta shqyrtojmë. Tekstet që njihen të jenë ose të Hamilton ose të Madison formojnë bashkësinë e trajtimeve ndërsa tekstet e veçanta dhe ato të përbashkëta do të formojnë bashkësinë test.

Në rastin kur tekstet individuale nuk janë shqyrtuar për qëllime krahasimi gjejmë se 4 nga 14 tekste të Madison janë përcaktuar si tekste të Hamilton, ndërsa vetëm 2 nga 52 tekste të Hamilton janë të klasifikuar gabim. Ky keq klasifikim i përgjithshëm në masën prej 9% është plotësisht i pranueshëm për krahasimin.

Të gjitha tekstet e paklasifikuara (keq klasifikuara) sipas zinxhirëve të Markovit kanë dalë të Madison. Si shtesë, artikujt e bashkuar psh nr 18 deri në 20 janë të klasifikuar respektivisht si të Madison, Hamilton dhe të Madison. Më vonë u vu re se fare pak nga 11 fjalët kyçe (funksion) shfaqeshin në tekst. Një teknikë thuajse njësoj me atë të përmendur më sipër që përdor një marrëveshje me bigramën e germave nuk do ta shfaqë këtë problem dhe premton të japi një rezultat të saktë .

Metoda jonë e përcakton artikullin 19 si të Hamilton megjithëse probabiliteti ndryshon vetëm me 0.0007 duke e krahasuar me diferencën mesatare prej 0.0048 për artikujt e padiskutueshëm.

1.6.4 Diskutime dhe përfundime

Shumë studime mbi njohjen e autorëve të huaj, tre prej të cilave u trajtuan në paragrafet e mesipërm, janë të limituara nga numri i vogël i teksteve që konsiderohen, gje e cila hap diskutime për përgjithësimin e tyre. Konkretisht, bashkësia e parë e të dhënave përbëhej nga 387 tekste nga 45 autorë, 74.42% e të cilave u klasifikuan në mënyrë korrekte. Në qoftë se në e trajtojmë si të rregullt një autor i cili është klasifikuar në treshen më të mirë koeficienti i suksesit do të rritet deri në 83.72% një rezultat mjaft i mirë. Khmelev et al (2001) raporton rezultate të njëjta me tekste të shkruara në rusisht.

Rezultatet e Khmelev et al (2001) u morën duke konsideruar korpuse të madhësive të ndryshme si korpusi prej qindra mijëra fjalësh nga arkiva e Projektit Gutenberg në korpusin me rreth dhjetë mijë në Guvat e Hijeve dhe korpusi me rreth 1000 fjalë nga Shkrimet Federalë.

Bashkësia e të dhënave e përdorur për zinxhirët e Markovit ndoshta mund të përshkruhet si mikroskopi linguistik, njësia është shumë e vogël për përfundime të shkëlqyera për të arritur në karakteristikat lidhëse të teksteve me autorët individualë. Krahasimi i matricave të kalimit mund ta lejojë studiuesin për të komentuar që Hamilton e përdor germën P të ndjekur nga një A më tepër se Madison por kjo gjë nuk hyn në stilin e interpretimit të tekstit.

Vargje të tilla germash ndoshta varen nga subjekti i tekstit. Rezultatet e improvizuara mund të merren duke mos marrë parasysh varësinë kontekst, fjalët dhe duke përformuar analizën vetëm në funksion të fjalës ose të fjalëve që shfaqen më shpesh. Një tjetër mundësi mund të jetë duke shqyrtuar kalimin midis pjesëve të fjalëve (ligjëratës) të përdorura, kështu që ndërhyjnë në strukturën sintaksore të tekstit duke shmangur çdo varësi në kontekst.

1.7 Aplikimi i metodës në tekstet e gjuhës shqipe

Në këtë paragraf do të bëhet një aplikimit i metodës në tekstet e gjuhë shqipe nëpërmjet programit të ndërtuar sipas përshkrimeve të paragrafit 1.5. Programi paraqitur në Paci et al (2011) llogariste koeficientet e entropisë vetëm duke u bazuar në matricat Markoviane të rendit të parë. Siç u përmend në paragrafin 1.4 programi u përmirësua duke bërë të mundur llogaritjen e koeficienteve të entropisë bazuar në matrica Markovi deri në rendin e gjashtë. Metoda është aplikuar në dy raste të veçanta, në 107 tekste të 13 autorëve të ndryshëm, ku 7 prej tyre janë romane të tre autorëve të ndryshëm dhe të tjerët janë artikuj të marra nga shtypi i përditshëm shqiptar nga 10 autorë të ndryshëm. Autoret janë shënuar Autori1, Autori2,..., Autori13. Tekstet përkatëse të autorit "*i"* janë shënuar Tekst.ij, Meqenëse alfabeti i gjuhës shqipe përbëhet edhe nga germat e përbëra metoda është aplikuar në dy raste studimi:

Rasti i parë:

Germat e përbëra nuk janë marrë në konsideratë, pra si alfabet është përdorur vargu i germave: A, B, C, Ç, D, E, Ë, F, G, H, I, J, K, L, M, N, O, P, Q, R, S, T, U, V, X, Y, , Z dhe karakteri hapësirë.

Rasti i dytë:

Germat e përbëra të gjuhës shqipe janë marrë në konsideratë, pra si alfabet është përdorur alfabeti i gjuhës shqipe A, B, C, Ç, D, DH, E, Ë, F, G, GJ, H, I, J, K, L, LL, M, N, NJ, O, P, Q, R, RR, S, SH, T, TH, V, U, X, XH, Y, Z, ZH dhe karakteri hapësire. Në këtë rast të gjitha germat përbëra janë zëvendësuar me numrat nga 1 deri në 9, pra si alfabet është përdorur vargu:

ABCÇD1EËFG2HIJKL3MN4OPQR5S6T7UVX8YZ 9

Fillimisht secili tekst është formatuar duke mos marrë në konsideratë llojin e shkrimit, germat e mëdha, shenjat e pikësimit si edhe karakteret e tjera jo germa të alfabetit të gjuhës shqipe. Me ndihmën e programit bëhet llogaritja e matricave të kalimit. Për secilin autor është llogaritur matrica e tij e kalimit si shumë e matricave të kalimit të teksteve të tij përkatëse. Në këtë mënyrë është populluar database i programit.

Për të përcaktuar autorësinë e një tekstit marrë nga database jonë, të cilin e supozoj si anonim (pa autorësi), pasi programi e përjashton këtë libër nga database, bëhet llogaritja e koeficienteve të entropisë sipas supozimit që ai tekst është shkruar nga secili prej autorëve të mësipërm. Autori që ka entropinë më të vogël do të jetë ai i cili ka shkruar tekstin e përzgjedhur si anonim. Metoda ka rezultuar e suksesshme në të dy rastet. Në të gjitha provat e kryera për secilin nga librat e përzgjedhur si anonim është përcaktuar autorësia e saktë. Rezultatet janë më të qarta në rastin e dytë sepse koeficienti i entropisë del edhe më i vogël se në rastin e parë, që do të thotë se sasia e informacionit të panjohur është zvogëluar. Gjithashtu rezultatet janë edhe më të qarta në qoftë se në tekst heqim të gjithë emrat e përveçëm të cilët nuk na japin ndonjë informacion të rëndësishëm për mënyrën e shkrimit të një autori. Metoda arrin të përcaktojë autorësinë e saktë edhe të një pjese të veçantë të shkëputur nga njëri prej teksteve.

1.7.1 Rezultatet në rastin e parë të studimit

Në këtë rast si alfabet për ndërtimin e matricës së kalimit është përdorur vargu i karaktereve: A, B, C, Ç, D, E, Ë, F, G, H, I, J, K, L, M, N, O, P, Q, R, S, T, U, V, X, Y, Z dhe karakteri hapësirë. Kjo nuk do të thotë që nuk konsiderohen germat e përbëra në tekst, por ato nuk konsiderohen si një klasë gramatikore.

Të gjithë tekstet janë të formatuar me germa të vogla dhe si file në PlainText. Fillimisht kemi populluar databasen e programit me 107 tekstet në gjuhën shqipe duke përcaktuar Alfabetin mbi të cilin ndërtohet matrica Markoviane. Në këtë rast Alfabeti është përcaktuar "ABCÇDEËFGHIJKLMNOPQRSTUVXY Z". Pasi është krijuar database me tekste është bërë përcaktimi i autorësisë për secilin tekst duke e supozuar atë si tekst pa autor. Në figurën 1.4 paraqitet faqja e programit në të cilën është llogaritur matrica markoviane për tekstin e katërt të autorit të parë "Tekst14". Zgjedhim tekstin në dosjen përkatëse pasi klikojmë në dritaren "Dosja", përcaktoj Alfabetin dhe më pas klikoj në butonin Llogarit për të llogaritur matricat Markoviane nga rendi i parë deri në rendin e gjashtë për ketë tekst. Më pas duke klikuar në butonin "Krahaso" programi kërkon rendin (figura 1.5) e zinxhirëve të Markovit për të cilin llogariten koeficientët e entropisë duke supozuar që teksti i përzgjedhur është shkruar nga secili autor, duke krahasuar matricën përkatëse të tekstit me matricat përkatëse të secilit autor. Në faqen "Rezultati shfaqen koeficientet e entropisë të renditura për secilin autor. Në figurën 1.6 paraqitet faqja "Rezultati" për tekstin "TEKST14". Sipas metodës, autori me entropinë më të vogël përcaktohet si autori i vërtetë i tekstit. Nga figura 1.6 vërejmë që autori me entropinë më të vogël rezulton të jetë autori i parë i cili është edhe autori i vërtetë i tekstit.

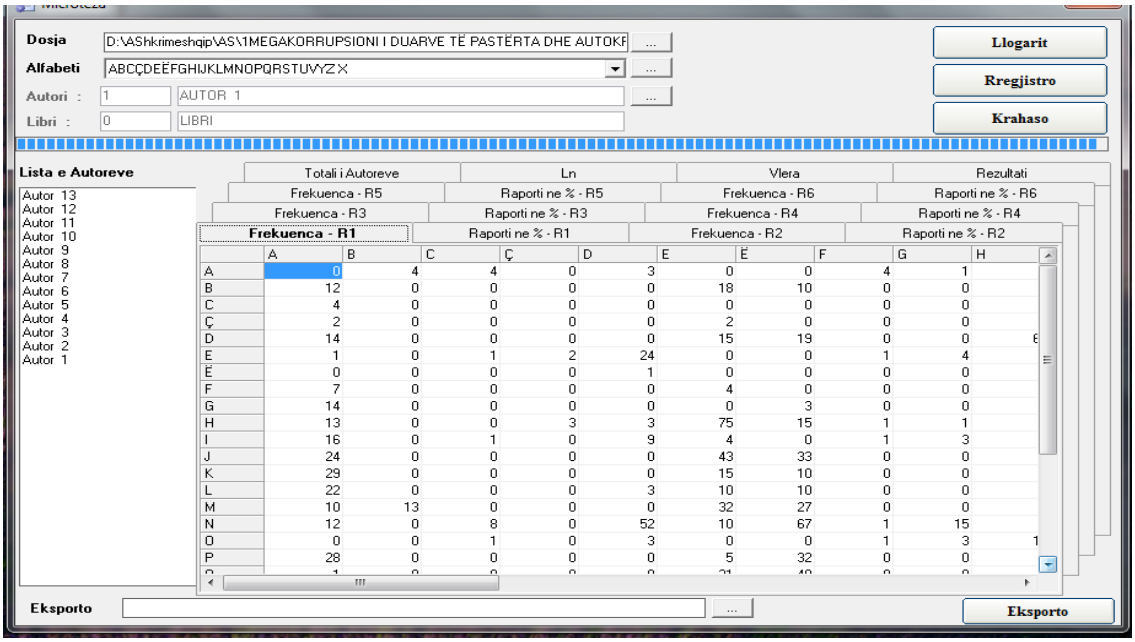

FIGURA 1. 4: LLOGARITJA E FREKUENCAVE TË SHFAQJES SË NJË GËRME PAS NJË GERME TJETËR, MBI BAZËN E SË CILËS NDËRTOHET MATRICA MARKOVIANE E RENDIT TË PARË.

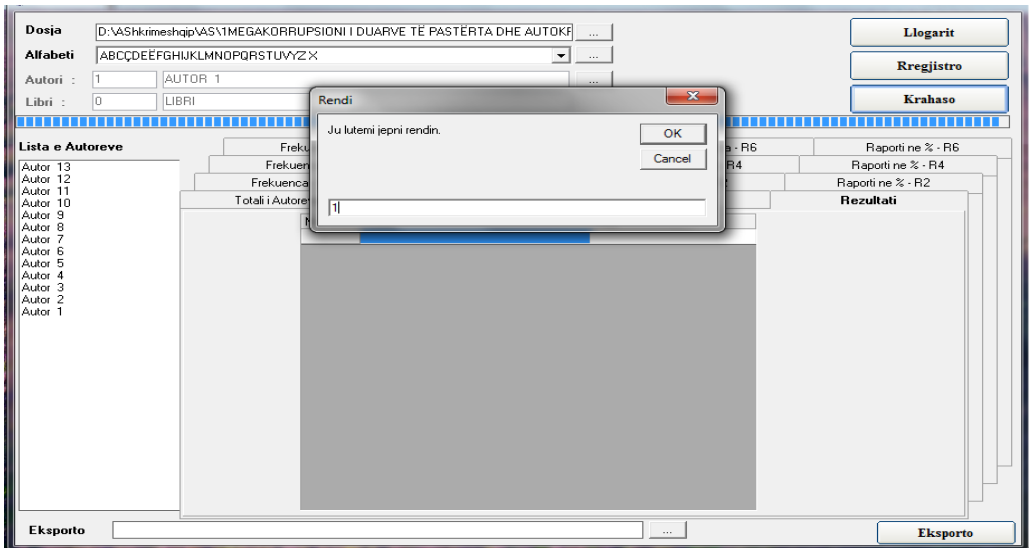

FIGURA 1. 5: CAKTIMI I RENDIT TË ZINXHIRIT TË MARKOVIT.

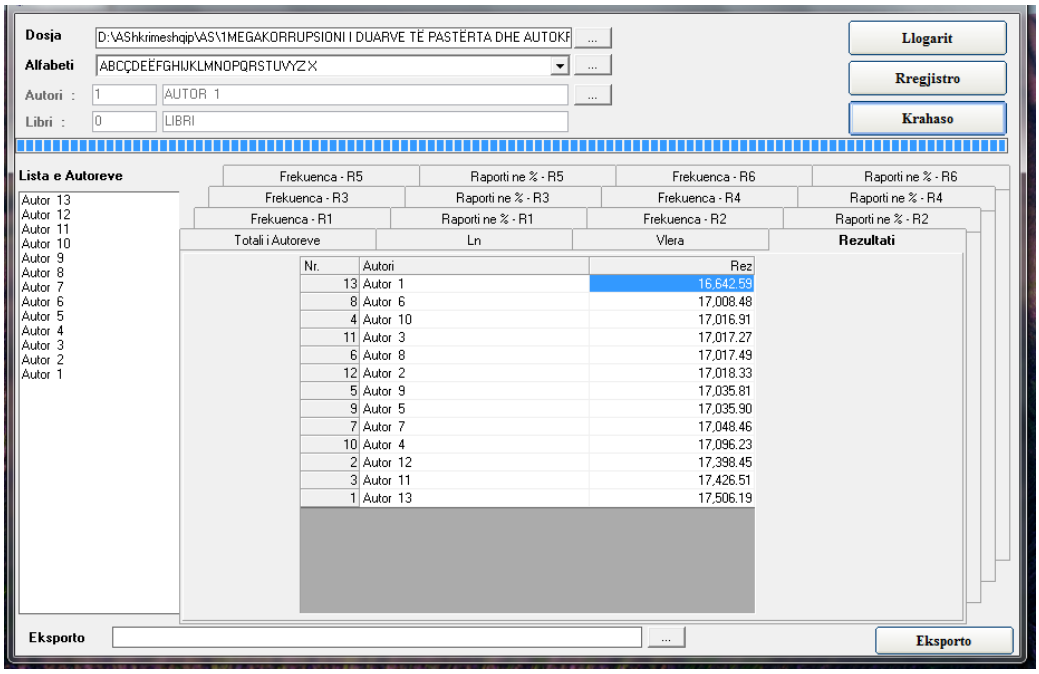

FIGURA 1. 6: REZULTATET E KOEFICIENTËVE TË ENTROPISË SIPAS SECILIT AUTOR, PËR TEKSTIN E ZGJEDHUR.

Në fund të faqes se programit është dritarja “Eksporto” e cila bën të mundur që të gjitha matricat dhe rezultatet që llogarit programi të ruhen si skedarë Excel. Të gjitha rezultatet e eksportuara në Excel për të gjitha testimet e bëra për secilin prej 106 teksteve shqip, nëpërmjet programit për të janë paraqitur në tabelat në Shtojca 1 në CD bashkëngjitur këtij punimi.

	TEKSTET E AUTORIT 1											TEKSTET E AUTORIT 2									
RENDI	1	2	3	4	5	6	7	8	9	10		1	2	3	4	5	6	7	8	9	10
1	1	1	1	1	1	1	1	1	1	1		4	2	2	2	7	7	2	9	7	2
2	1	1	1	1	8	1	1	1	1	1		4	2	2	2	2	7	2	2	1	2
3	1	1	1	1	8	1	1	7	7	10		7	2	2	2	2	7	1	2	1	2
4	1	1	3	1	8	1	1	7	7	10		7	2	2	2	2	7	1	2	7	2
5	3	7	3	1	4	1	1	7	7	10		7	6	2	2	4	7	1	2	10	5
6	3	7	3	1	4	1	1	7	7	10		7	6	2	2	4	7	1	2	10	5
	TEKSTET E AUTORIT 3											TEKSTET E AUTORIT 4									
RENDI	1	2	3	4	5	6	7	8	9	10		1	2	3	4	5	6	7	8	9	10
1	3	1	3	3	3	3	3	3	3	3		6	4	4	4	4	4	4	4	4	5
2	3	3	3	3	3	3	3	3	3	3		4	4	4	4	4	4	4	4	4	4
3	3	3	3	3	3	3	3	3	3	3		8	4	10	4	4	4	4	4	4	4
4	3	4	3	3	3	3	3	3	3	9		8	4	10	4	4	4	10	4	4	4
5	3	4	3	3	3	3	3	3	3	9		8	4	10	4	4	4	10	4	4	4
6	3	4	3	3	3	3	3	3	3	9		8	4	10	4	4	4	10	4	4	4
	TEKSTET E AUTORIT 5											TEKSTET E AUTORIT 6									
RENDI	1	2	3	4	5	6	7	8	9	10		1	2	3	4	5	6	7	8	9	10
1	5	5	5	5	5	5	5	5	5	5		6	6	6	4	6	6	6	6	6	6
2	5	5	5	5	5	5	5	5	5	5		6	6	6	6	6	6	6	6	3	6
3	5	5	5	5	5	5	5	5	5	10		9	10	6	6	6	6	6	6	10	3
4	5	5	5	5	5	5	5	5	5	10		9	3	6	6	6	10	10	6	10	3
5	5	8	5	8	5	5	5	5	5	1		9	10	6	6	6	10	10	6	10	10
6	5	8	5	8	5	5	5	5	5	1		9	10	6	6	10	10	10	6	10	10
	TEKSTET E AUTORIT 7											TEKSTET E AUTORIT 8									
RENDI	1	2	3	4	5	6	7	8	9	10		1	2	3	4	5	6	7	8	9	10
1	7	7	7	7	7	7	7	7	7	7		3	8	8	8	4	8	8	4	8	8
2	5	7	7	7	7	7	7	7	2	7		3	8	8	8	8	8	4	8	8	8
3	5	7	7	7	7	7	7	7	2	7		3	2	8	8	8	8	8	8	8	8
4	4	7	7	7	7	7	7	7	2	10		3	2	8	8	8	7	8	8	8	3
5	4	7	7	7	7	7	7	7	2	10		3	2	8	8	8	7	8	8	8	3
6	4	7	7	7	7	7	7	7	2	10		3	4	8	8	8	7	8	8	8	5

	TEKSTET E AUTORIT 9											TEKSTET E AUTORIT 10									
RENDI	1	2	3	4	5	6	7	8	9	10		1	2	3	4	5	6	7	8	9	10
1	9	9	9	4	9	9	9	9	9	9		10	10	10	10	10	10	8	10	10	10
2	9	9	9	9	9	9	9	9	9	9		10	10	10	10	10	10	10	10	10	10
3	9	9	9	9	9	9	9	9	9	9		10	10	10	2	10	10	10	10	10	10
4	9	9	9	9	9	9	5	9	9	9		10	10	10	2	10	10	10	10	10	10
5	9	9	9	9	9	9	5	9	9	9		10	10	8	2	10	10	4	10	10	8
6	9	9	9	9	9	9	5	9	9	9		10	10	8	2	10	10	11	10	10	8

TABELA 1. 2: PËRCAKTIMI I AUTORIT PËR SECILIN ARTIKULL NË RASTIN E PARE TË STUDIMIT DUKE U BAZUAR NË ZINXHIRET E MARKOVIT DERI NË RENDIN E GJASHTË .

	TEKSTET E AUTORIT 11				TEKSTET E AUTORIT 12		TEKSTET E AUTORIT 13
RENDI	1	2	3	4	5	6	7
1	11	11	11	11	12	12	13
2	11	11	11	11	12	12	13
3	11	11	11	11	12	12	13
4	11	11	11	11	12	12	13
5	11	11	11	11	12	12	13
6	11	11	11	11	12	12	13

TABELA 1. 3: PËRCAKTIMI I AUTORIT PËR SECILIN ROMAN NË RASTIN E PARE TË STUDIMIT DUKE U BAZUAR NË ZINXHIRET E MARKOVIT DERI NË RENDIN E GJASHTË .

Nga rezultatet e tabelës 1.3 vërejmë se identifikimi i autorit për tekstet e mëdha me shumë fjalë, në rastin tonë romanet, është shumë i saktë. Kur kontrollohet autorësia për tekstet që janë artikuj gazetash autorët e romaneve gjithmonë renditen të fundit. Pra në metodë ndikon edhe madhësia e tekstit.

Nga tabelat 1.2 dhe 1.3 përmbledhim rezultatet për tekstet të cilëve ju është identifikuar autori i saktë në tabelën 1.4.

RENDI	NUMRI I TEKSTEVE të IDENTIFIKUAR SAKTE	PROBABILITETI I SUKSESIT
1	93	0.869159
2	98	0.915888
3	86	0.803738
4	77	0.719626
5	66	0.616822
6	65	0.607477

TABELA 1. 4: PROBABILITETET E SUKSESIT NË RASTIN E PARË TË STUDIMIT PËR SECILIN MODEL.

Nga tabela 1.4 vërehet se probabiliteti i suksesit për identifikimin autorit me modelin e zinxhirëve të Markovit të rendit të parë, është me i madh krahasuar me rezultatet në punimet Salillari et al (UNIEL2012) dhe Salillari et al (ICEOS2012), ku aplikimi është realizuar në një numër më të vogël tekstesh. Pra me rritjen e numrit te teksteve metoda jep rezultate më të mira. Identifikimi i autorit për zinxhirët e Markovit të rendit të dytë është bërë me saktësinë më të madhe, dhe pas rendit të tretë kemi një rënie të probabilitetit të identifikimit të autorësisë deri në 0.607 ndryshe nga rezultatet qe pritej te merrnim. Për te kuptuar arsyen pse saktësia e identifikimit të autorësisë së teksteve u arrit me modelin e Zinxhirit të Markovit të rendit të dytë duhet të bëhet një analizë statistikore dhe të studiohet shpërndarja probabilitare e fjalëve në këto tekste. Analiza statistikore e teksteve do të trajtohet më poshtë në kapitullin e dytë. Si rezultat i ndërtimit te shpërndarjes probabilitare të fjalëve në kapitullit 2 jepet se fjalët më të përdorura në tekstet shqip janë "për", "dhe", "një", "nuk", "nga" ku secila prej tyre ka vetëm tre germa sipas Alfabetit në ketë rast studimi, kështu që modeli i zinxhirëve të Markovit të rendit të dytë i cili bazohet në probabilitetet e kalimit me dy hapa nga një germë në tjetrën është modeli më i mirë për identifikimin e autorësisë se teksteve shqip.

1.7.2 Rezultatet në rastin e dytë të studimit

Në këtë rast germat e përbëra të gjuhës shqipe janë marrë në konsideratë, pra si alfabet është përdorur alfabeti i plotë i gjuhës shqipe A, B, C, Ç, D, DH, E, Ë, F, G, GJ, H, I, J, K, L, LL, M, N, NJ, O, P, Q, R, RR, S, SH, T, TH, V, U, X, XH, Y, Z, ZH dhe karakteri hapësirë. Gjatë llogaritjeve të gjitha germat përbëra janë zëvendësuar me numrat nga 1 deri në 9, pra si alfabet është përdorur vargu:

A, B, C, Ç, D, 1, E, Ë, F, G, 2, H, I, J, K, L, 3, M, N, 4, O, P, Q, R, 5, S, 6, T, 7, U, V, X, 8, Y, Z, 9

Është treguar kujdes në formatimin e teksteve duke i formatuar ata më parë nga të gjithë numrat e çfarëdoshëm në përmbajtje të tekstit si edhe numrat e faqeve. Më pas në secilin tekst të gjitha germat e përbëra janë zëvendësuar me numrat 1 deri në 9. Të gjithë tekstet janë formatuar në germa të vogla dhe janë ruajtur si PlainText, të cilëve u është vendosur 2 përpara titullit për ti dalluar nga tekstet e formatuar për rastin e parë të studimit. Fillimisht kemi populluar databasen e programit me 107 tekstet në gjuhën shqipe duke përcaktuar Alfabetin mbi të cilin ndërtohet matrica Markoviane. Në ketë rast Alfabeti është përcaktuar "ABCÇD1EËFG2HIJKL3MN4OPQR5S6T7UVX8YZ 9". Pasi është krijuar database me tekste, është realizuar përcaktimi i autorësisë për secilin tekst duke e supozuar atë si tekst pa autor. Në figurën 1.7 paraqitet faqja e programit në të cilën është llogaritur matrica markoviane për tekstin e katërt të autorit të parë "Tekst14". Ndiqet e njëjta procedure si në rastin e parë të studimit duke përcaktuar Alfabetin e sipër përmendur dhe më pas klikojmë në butonin Llogarit për të llogaritur matricat Markoviane nga rendi i pare deri në rendin e gjashtë për ketë tekst. Në figurën 1.8 paraqitet faqja "Rezultati" për tekstin "TEKST14" për të cilin entropia me e vogël ka rezultuar përsëri ajo e "Autori1" i cili është edhe autori i vërtete i tekstit. Por vërejmë se thuajse të gjithë koeficientet e entropisë janë më të vegjël krahasuar me ata të paraqitur në figurën 1.6. Kjo gjë vërehet thuajse në të gjitha rezultatet e marra në ketë rast studimi. Në këtë rast rezultatet janë më të qarta se në rastin e parë sepse koeficienti i entropisë përkatëse del edhe më i vogël se në rastin e parë, që do të thotë se sasia e informacionit të panjohur është zvogëluar.

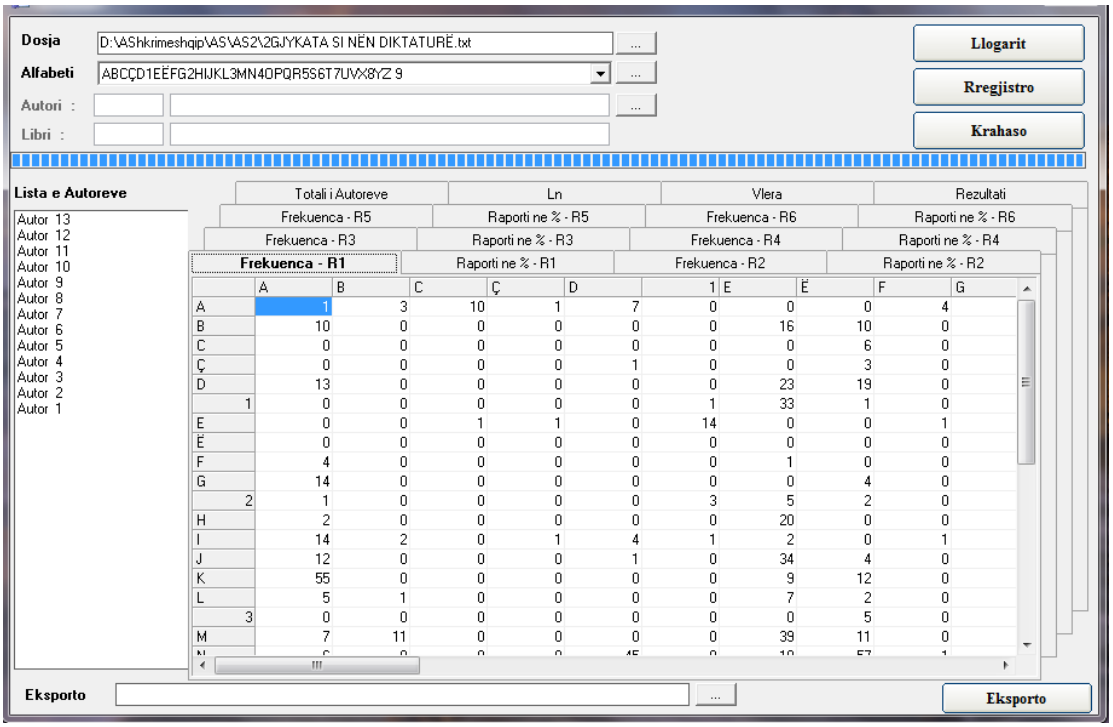

FIGURA 1. 7: MATRICA E FREKUENCAVE PER TEKSTIN TEKST14.

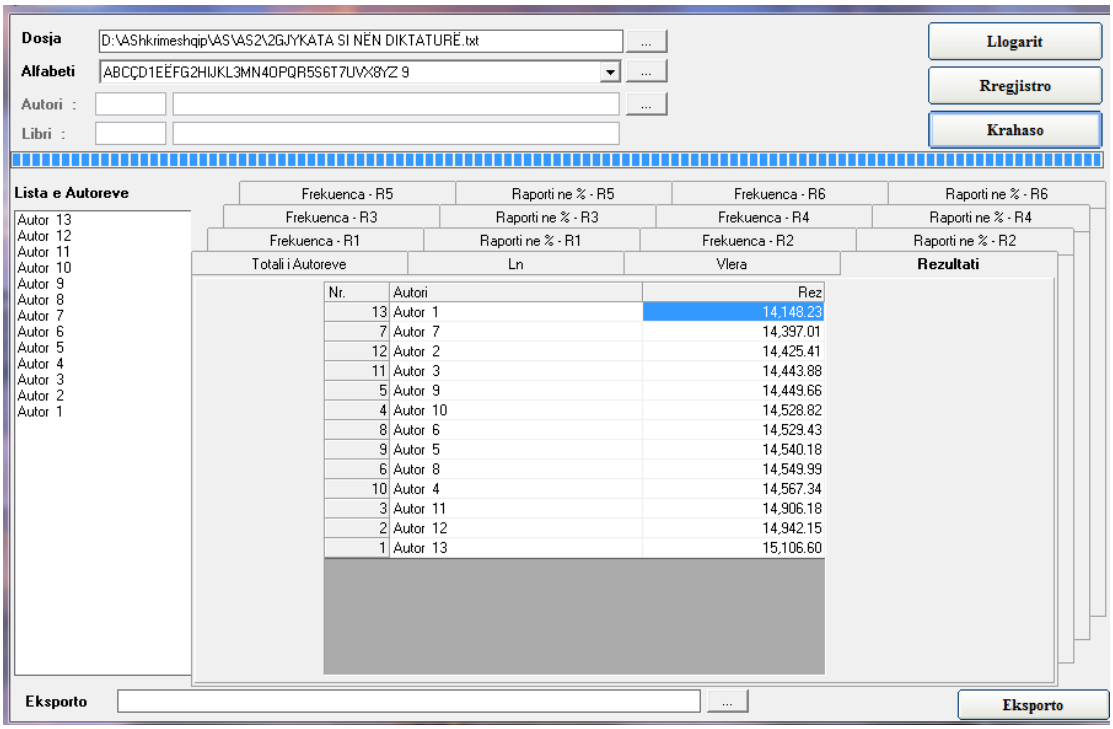

FIGURA 1. 8: REZULTATET PER ZINXIRIN E MARKOVIT TË RENDIT TË PARE PER TEKST14.

	TEKSTET E AUTORIT 1											TEKSTET E AUTORIT 2									
RENDI	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**		**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
1	1	1	1	1	1	1	1	1	1	1		2	4	2	2	7	9	1	9	7	4
2	1	1	1	1	8	1	1	1	1	1		2	7	2	2	2	2	2	2	7	2
3	1	1	1	1	8	1	1	1	1	1		2	7	2	2	2	7	1	2	7	2
4	1	7	1	1	8	1	1	7	7	10		2	7	2	2	2	7	1	2	7	2
5	2	7	1	1	8	1	1	7	7	10		2	7	2	2	2	7	1	2	7	2
6	10	7	1	1	8	1	1	7	7	10		2	7	2	2	2	7	1	2	7	10
	TEKSTET E AUTORIT 3											**TEKSTET E AUTORIT 4**									
RENDI	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**		**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
1	3	3	3	3	3	3	3	3	3	3		7	4	4	4	4	4	4	4	4	5
2	3	3	3	3	3	3	3	3	3	3		4	4	4	4	4	4	4	4	4	4
3	3	4	3	3	3	3	3	3	3	3		10	4	4	4	4	4	4	4	4	4
4	3	4	3	3	3	3	3	3	3	3		10	4	10	4	4	4	10	4	4	4
5	3	4	3	3	3	3	3	3	3	3		10	4	10	4	4	4	10	4	4	4
6	3	4	3	3	3	10	3	3	3	3		10	4	10	4	4	4	10	4	4	4

	TEKSTET E AUTORIT 5											TEKSTET E AUTORIT 6									
RENDI	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**		**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
1	5	5	5	4	3	5	5	5	5	5		6	6	4	4	6	6	6	6	6	6
2	5	5	5	5	5	5	5	5	5	5		6	6	6	6	6	6	6	6	6	6
3	5	5	5	5	5	5	5	5	5	10		9	10	6	6	6	6	6	6	6	9
4	5	5	5	5	5	5	5	5	5	10		9	10	6	6	10	6	4	6	10	9
5	5	4	5	5	5	5	5	5	5	10		9	5	6	6	10	6	4	6	10	9
6	5	4	5	8	5	5	5	5	5	10		9	5	6	6	10	7	4	6	10	9
	TEKSTET E AUTORIT 7											**TEKSTET E AUTORIT 8**									
RENDI	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**		**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
1	7	7	7	7	7	7	7	7	7	7		3	8	8	8	4	8	8	9	8	8
2	7	7	7	7	7	7	7	7	2	7		3	8	8	8	8	8	8	8	8	8
3	7	7	7	7	7	7	7	7	2	7		3	8	8	8	8	7	8	8	8	8
4	4	7	7	7	7	7	7	7	2	7		3	8	8	8	8	7	8	8	8	5
5	4	7	7	7	7	7	7	7	2	7		3	8	8	8	8	7	8	8	8	5
6	4	7	7	7	7	7	7	7	2	10		3	4	8	8	8	7	8	8	8	5
	TEKSTET E AUTORIT 9											**TEKSTET E AUTORIT 10**									
RENDI	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**		**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
1	9	9	9	4	9	9	9	9	9	9		10	9	10	10	10	10	8	10	10	10
2	9	9	9	9	9	9	9	9	9	9		10	10	10	10	10	10	10	10	10	10
3	9	9	9	9	9	9	9	9	9	9		10	10	10	2	10	10	10	10	10	10
4	9	9	9	9	9	9	9	9	9	9		10	10	10	2	10	10	10	10	10	10
5	9	9	9	9	9	9	5	9	9	9		10	10	10	10	10	10	10	10	10	10
6	9	9	9	9	9	9	10	9	9	9		10	10	10	10	10	10	10	10	10	10

TABELA 1. 5: REZULTATET E IDENTIFIKIMIT TË AUTOREVE TË ARTIKUJVE NË RASTIN E DYTE TË STUDIMIT.

RENDI	TEKSTET E AUTORIT 11				TEKSTET E AUTORIT 12		TEKSTET E AUTORIT 13
	1	2	3	4	5	6	7
1	11	11	11	11	12	12	13
2	11	11	11	11	12	12	13
3	11	11	11	11	12	12	13
4	11	11	11	11	12	12	13
5	11	11	11	11	12	12	13
6	11	11	11	11	12	12	13

TABELA 1. 6 REZULTATET E IDENTIFIKIMIT TË AUTOREVE TË ROMANEVE NË RASTIN E DYTE TË STUDIMIT.

Duke përmbledhur rezultatet e tabelave 5 dhe 6 përftojmë rezultatet në tabelën 7 si me poshtë:

RENDI	NUMRI I TEKSTEVE të IDENTIFIKUAR SAKTE	PROBABILITETI I SUKSESIT
1	90	0.841121
2	102	0.953271
3	92	0.859813
4	81	0.757009
5	79	0.738318
6	73	0.682243

TABELA 1. 7: PROBABILITETET E IDENTIFIKIMIT TË AUTORIT TË TEKSTEVE NË RASTIN E DYTE TË STUDIMIT

Nga tabela 1.7 vërejmë se identifikimi i autorësisë së teksteve bëhet me saktësi më të madhe në rastin e zinxhirëve të Markovit të rendit të dytë dhe të rendit të tretë.

Krahasuar me rastin e parë të studimit konstatojmë se kemi një rënie të probabilitetit për identifikimin e autorit në rastin e Zinxhirëve të Markovit të rendit të parë nga 0.869 në 0.84 dhe kemi një rritje të ndjeshme të këtij probabiliteti për të gjithë rendet e tjera të

zinxhirëve të Markovit. Kjo ndodh për arsye se duke konsideruar shkronjat e përbëra si një germë të vetme kemi ndryshim të numrit të fjalëve me 2 shkronja në tekste. Në këtë rast studimi vërehet një zvogëlim i koeficienteve të entropisë e cila do të thotë se në këtë rast studimi kemi më pak informacion të panjohur për tekstet. Saktësia e identifikimit të autorësisë së teksteve është bërë me probabilitetin më të lartë 0.953 në rastin e zinxhirëve të Markovit të rendit të dytë i cila është një rezultat shumë i mirë dhe përcakton edhe modelin më të mirë të zinxhirëve të Markovit për identifikimin e autorit në tekstet shqip.

Vërejmë se thuajse në të gjithë rastet e krahasimit entropitë janë më të vogla krahasuar me rastin e parë të trajtuar në paragrafin më sipër. Kjo tregon se duke konsiderua germat e përbëra si një klasë gramatikore ato na japin informacion të vlefshëm për mënyrën e të shkruarit të shkrimtarëve.

Duhet përmendur që rezultatet e marra për modelet e zinxhirëve të Markovit në gjuhën shqipe janë shumë të mira krahasuar me rezultatet e përftuara nga Khmelev et al (2001) të cilët konsiderojnë rezultat shumë të mirë rritjen e koeficientin e suksesit nga 74.42% deri në 83.72% ndërkohë që duke konsideruar modelet nga rendi i parë në rendin e katërt të Zinxhirëve të Markovit për rastin dytë të studimit vërehet një rritje të koeficientit të suksesesit nga 75.7% në 95.3%

1.8 Përfundime

Metoda jep rezultate shumë të mira kur aplikohet në një numër të madh tekstesh kjo për arsye se në një numër të madh tekstesh kemi informacion të mjaftueshëm për frekuencat e karaktereve në tekste të autorëve.

Metoda aplikohet në tekstet e gjuhës Shqipe në dy rastet e sipër përmendura. Në rastin e dytë rezultatet janë më të qarta se në rastin e parë sepse koeficienti i entropisë del edhe më i vogël se në rastin e parë, që do të thotë se sasia e informacionit të panjohur është zvogëluar. Metoda rrit saktësinë e identifikimit të autorësisë në tekstet shqip kur vlerësohet entropia për matrica Markoviane të rendeve të larta. Modeli më i mirë i zinxhirëve të Markovit për identifikimin e autorit në tekstet shqip është modeli i zinxhirëve të Markovit të rendit të dytë në rastin kur konsiderojmë germat e përbëra të alfabetit shqip. Për tekstet në gjuhën shqipe saktësia e identifikimit të autorësisë për këtë model rezultoi 95.3 %.

Rezultatet e përftuar për të gjashtë modelet e zinxhirëve të markovit të ndërtuar për të dy rastet e studimit për tekstet në gjuhën shqipe janë rezultate shumë me të mira krahasuar me rezultatet e marra nga modeli i zinxhirit të Markovit të rendit të pare për tekstet në gjuhën Angleze dhe Ruse nga Khmelev et al (2001).

Gjithashtu rezultatet janë edhe më të qarta në qoftë se në tekst heqim të gjithë emrat e përveçëm të cilët nuk na japin ndonjë informacion të rëndësishëm për mënyrën e shkrimit të një autori.

Metoda arrin të përcaktojë autorësinë e saktë edhe të një pjese të veçantë të shkëputur nga njëri prej teksteve.

Kapitulli 2

Analize statistikore e teksteve në gjuhën shqipe

2.1 Hyrje

Në këtë kapitull është bërë përpunimi i 107 teksteve në gjuhën shqipe duke krijuar një korpus me 4666163 fjalë, për të bërë analizë statistikore të elementeve gjuhësor, duke prezantuar disa paketa të programit R.

2.2 Ligji Zipf

Në qoftë se numërojmë sa herë shfaqet një germë apo një fjale në një korpus tekstesh të një gjuhe dhe i rendisim këto frekuenca atëherë shqyrtohet lidhja midis frekuencës *"f"* të germës (fjalës) dhe pozicionit *"r" të* saj në renditje (rendi). Ky fakt është dokumentuar për shpërndarjen e fjalëve të gjuhës nga Zipf(1949). Ligji Zipf pohon se frekuenca e germës është e përpjesshme me *"1/r"* pra $f(r) \propto \frac{1}{r^{\alpha}}$ ku $\alpha \approx 1$ ose me fjalë të tjera ekziston një konstante *k* e tille qe $f * r = k$. Përgjithësimi i ligjit Zipf ka trajtën:

$$f(r) = \frac{C}{r^{\alpha}}$$

ku α është një parametër pozitiv me vlera shumë afër 1 dhe C një konstante.

Për të arritur një përshtatje më të mirë të shpërndarjes së fjalëve të një gjuhe Mandelbrot(1962) nxorri një përgjithësim tjeter të këtij ligji si me poshtë:

$$f = P(r + \rho)^{-\alpha} \quad ose \quad logf = logP - \alpha log(r + \rho)$$

ku P një konstante, $\alpha \approx 1$, $\rho \approx 2.7$ janë parametra të një teksti të cilat se bashku matin pasurinë e përdorimit të fjalës në tekst.

Ligji zipf dhe përgjithësime të tij është ligj të cilit i binden shumica e gjuhëve të shkruara dhe të folura. Në paragrafët më poshtë ndërtohet shpërndarja e germave dhe të fjalëve në gjuhën e shkruar shqipe për korpusin me 107 tekste shqip.

2.3 Analiza e grupimeve

Qëllimi thelbësor i analizës së grupimeve (*clustering*) është ndarja e të dhënave në grupime (nënbashkësi) në mënyrë të tillë që të dhënat brenda të njëjtit grupim të jenë sa më të ngjashme me njëra-tjetrën dhe sa më të ndryshme nga të dhënat në grupime të tjera. Koncepti i ngjashmërisë shprehet në mënyrë sasiore me anë të funksioneve distance. në gjuhën e shkruar analiza e grupimeve është një metode klasifikimi e cila grupon fjalët për nga ngjashmëria e shpërndarjes se tyre në tekste. Për trajtimin e problemit të analizës së grupimeve ekziston një larmi algoritmesh një kategori e të cilëve aplikohet për gjuhën e shkruar janë algoritmet hierakikë.

Algoritmet hierakikë të analizës së grupimeve synojnë të ndërtojnë një hieraki grupimesh në një strukturë në formën e një peme (të quajtur "dendrogramë"). Për të gjeneruar pemën e grupimeve, algoritmet hierarkikë mund të veprojnë në mënyrë ndarëse (nga lart poshtë) ose në mënyrë bashkuese (nga poshtë lart) dhe për të kryer ndarjet/bashkimet bazohen në kritere të cilat shfrytëzojnë funksionet distancë siç është përmendur nga Alpaydin (2010).

Një funksion distancë është një funksion i formës $d{:}X\times X\to R$ i cili gëzon vetitë:

(1) $d(x,y)\geq 0, \forall x,y\in X$

(2) $d(x,y)=d(y,x), \forall x,y\in X$

(3) $d(x,y)=0$ atëherë dhe vetëm atëherë kur $x=y\ \forall x,y\in X$

(4) $d(x,y)\leq d(x,z)+d(z,y), \forall x,y,z\in X$ (mosbarazimi i trekëndëshit)

Le të shënojmë bashkësinë e të dhënave $X=\{x_1,x_2,\dots,x_n\}$. Gjithashtu le të jetë përzgjedhur një funksion $d(x,y)$ për të shprehur distancën midis dy elementëve individualë x dhe y, si dhe një funksion $D(G_i,G_j)$ për të shprehur distancën midis grupimeve G_i dhe G_j të bashkësisë momentale të grupimeve $\Gamma=\{G1,G2,\dots,Gk\}$.
Llogariten distancat ndërmjet grupimeve, që momentalisht janë:

$$D(G_i,G_j)=d(x_i,x_j)$$

meqë çdo grupim përbëhet nga një e dhënë. Në bashkësinë e grupimeve Γ gjejmë çiftin e grupimeve (G_u,G_v) për të cilët distanca midis tyre është më e vogla, pra
$D(Gu,Gv)=\min_{P,Q\in\Gamma} D(P,Q)$.
I bashkojmë grupimet Gu dhe Gv në një grupim Gw, pra $Gw=Gu\cup Gv$. Largohen Gu dhe Gv nga Γ , dhe shtohet Gw në Γ.
Ekzistojnë disa variante të algoritmit hierarkik bashkues, në varësi të mënyrës si shprehet distanca $D(G_i,G_j)$ midis grupimeve. Disa nga variantet kryesore janë algoritmi hierakik i lidhjes njëfishe (minimale), algoritmi hierakik i lidhjes së plotë (maksimale) dhe algoritmi hierarkik i lidhjes mesatare.

Në gjuhën e shkruar analiza e grupimeve behet me metodën e algoritmit hierakik të lidhjes së plotë.

Algoritmi hierarkik i lidhjes së plotë, distancën midis dy grupimesh e shpreh si distanca më e madhe që ekziston midis ndonjë elementi të grupimit të parë dhe ndonjë elementi të grupimit të dytë, pra distanca midis dy grupimesh G_p dhe G_q shprehet si:

$$D(Gu, Gv) = \max_{x \in Gu, y \in Gv} d(x, y)$$

2.4 Gjuha R dhe paketat për përpunimin e teksteve

Programi i gjuhës R është një program i zhvilluar në mënyrë specifike për analizën statistikore të të dhënave. Ky program përdor një gjuhë programimi e cila rrjedh nga gjuha S. Programi përshtatet me të gjitha sistemet operative. Ky program është i lehtë në përdorim, duke qënë se nuk kërkohen njohuri të avancuara mbi gjuhët e programimit për të punuar me të. Programi, gjithashtu, është falas, pra mund të shkarkohet direkt nga faqja e tij e internetit pa qënë nevoja të paguhet për të. Shkarkimi i tij mund të bëhet në: http://www.r-project.org/. Popullariteti i këtij programi lidhet me shumëllojshmërinë e paketave në dispozicion, të cilat i zgjerojnë më tej kapacitetet e programit, duke ofruar funksione shtesë përveç atyre që ekzistojnë tashmë në program. Këto paketa janë përmbledhur në sistem të quajtur CRAN, i cili mund të gjendet në: http://cran.r-project.org/. Në këtë faqe është i mundur të bëhet edhe shkarkimi i vetë programit.

Siç e përmendëm dhe më sipër në dispozicion të programit janë disa lloje paketash, në të cilat përmbahen shumë funksione shtesë të cilat nuk përfshihen në versionin e sapo instaluar të programit. Kjo vjen si rrjedhojë e faktit se këto funksione janë më specifike, përdoren për probleme më specifike.

Programi i gjuhës R perveçse një program për analizën statistikore të të dhënave ai përmban edhe paketa për përpunimin e teksteve (Text mining). Për përpunimin e teksteve në gjuhën shqipe dhe për të ndërtuar një shpërndarje të fjalëve dhe germave në gjuhën shqipe është përdorur pikërisht programi R. Paketat e nevojshme që implementohen në R për përpunimin e teksteve janë "tm", "SnowballCC", "RColorBrewer", "ggplot2", "wordcloud", "biclust", "cluster", "igraph", "fpc". “zipfR”. Zhao (2012).

2.5 Analiza statistikore e teksteve shqip në R

Fillimisht kemi importuar tekstet në R duke krijuar tre direktori “tekste”, “artikuj” dhe “romane” ku direktoria e parë është bashkim i dy direktorive të tjera. Kjo është bërë për qëllim të studimit veçmas romanet dhe artikujt meqenëse janë tekste që kanë madhësi të ndryshme. Me pas nëpërmjet funksionit Corpus() kemi krijuar tre korpuse përkatëse i pari me 107 tekste në gjuhën shqipe, i dyti me 100 artikuj të 10 gazetareve të ndryshëm shqiptare dhe i treti me 7 romane të 3 shkrimtareve të njohur shqiptare si me poshtë:

```
> tekste<-file.path("C:","teksteshqip")
> tekste
```

```
[1] "C:/teksteshqip"
> dir(tekste)

> artikuj<-file.path("C:","texts")
> artikuj
[1] "C:/texts"
> dir(artikuj)
> romane<-file.path("C:","romaneshqip")
> romane
[1] "C:/romaneshqip"
> dir(romane)

> library(tm)
> shqip<-Corpus(DirSource(tekste))
> shqip
<<VCorpus>>
Metadata:  corpus specifi c: 0, document level (indexed): 0
Content:  documents: 107
> dok<-Corpus(DirSource(artikuj))
> dok
<<VCorpus>>
Metadata:  corpus specific: 0, document level (indexed): 0
Content:  documents: 100

> dokr<-Corpus(DirSource(romane))
> dokr
<<VCorpus>>
Metadata:  corpus specific: 0, document level (indexed): 0
Content:  documents: 7
```

Për të përpunuar tekstet është e nevojshme të përjashtojmë shenjat e pikësimit, numrat, germat e mëdha si dhe hapësirat e panevojshme midis fjaleve që ndodhen në këto korpuse. Duhet qe tekstet e korpusit të lexohen si file në formatin PlainTextDocument. Këto i realizojmë me ndihmën e funksionit tm_map(). Më pas për korpuset përkatëse krijojmë një matricë termash të saj nepërmjet funksionit DocumentTermMatrix(). Më poshtë po japim përpunimin e teksteve në korpusin "shqip":

```
> shqip<-tm_map(shqip,removePunctuation)
> shqip<-tm_map(shqip,removeNumbers)
> shqip<-tm_map(shqip,tolower)
> library(SnowballC)
> shqip <- tm_map(shqip, stripWhitespace)
> shqip <- tm_map(dokr, stemDocument)
> shqip<-tm_map(shqip, PlainTextDocument)
```

```
> dtmsh <- DocumentTermMatrix(shqip)
> dtmsh
<<DocumentTermMatrix (documents: 107, terms: 43609)>>
Non-/sparse entries: 112370/4553793
Sparsity           : 98%
Maximal term length: 53
Weighting          : term frequency (tf)
```

Nga përshkrimi i matricës së termave të matricës për korpusin "shqip" vërejmë se kemi një përqindje shumë të lartë 98% të fjalëve me frekuence të vogël. Pra në ketë korpus kemi gjithsej 4666163 fjalë nga të cilat 112370 kanë frekuencë të lartë.

Në të njëjtën mënyrë realizojmë përpunimet e teksteve edhe në dy korpuset e tjerë si më poshtë:

```
> dok<-tm_map(dok,removePunctuation)
> inspect(dok[2])
<<VCorpus>>
Metadata:  corpus specific: 0, document level (indexed): 0
Content:  documents: 1

[[1]]
<<PlainTextDocument>>
Metadata:  7
Content:  chars: 6533
> dok<-tm_map(dok,removeNumbers)
> inspect(dok[2])
<<VCorpus>>
Metadata:  corpus specific: 0, document level (indexed): 0
Content:  documents: 1
[[1]]
<<PlainTextDocument>>
Metadata:  7
Content:  chars: 6516
> dok<-tm_map(dok,tolower)
> library(SnowballC)
> dok <- tm_map(dok, stripWhitespace)
> dok <- tm_map(dok, PlainTextDocument)
> dtm <- DocumentTermMatrix(dok)
> dtm
<<DocumentTermMatrix (documents: 100, terms: 19246)>>
Non-/sparse entries: 52851/1871749
Sparsity           : 97%
Maximal term length: 53
```

```
Weighting          : term frequency (tf)
```

Nga përshkrimi i matrices vërejmë se 97% e termave të korpusit jane të rralla pra me një frekuence të ulet.

```
> dokr<-tm_map(dokr,removePunctuation)
> dokr<-tm_map(dokr,removeNumbers)
> dokr<-tm_map(dokr,tolower)
> library(SnowballC)
> dokr <- tm_map(dokr, stemDocument)
> dokr <- tm_map(dokr, stripWhitespace)
> dokr <- tm_map(dokr, PlainTextDocument)
> dtmr <- DocumentTermMatrix(dokr)
> dtmr
<<DocumentTermMatrix (documents: 7, terms: 33099)>>
Non-/sparse entries: 59831/171862
Sparsity           : 74%
Maximal term length: 36
Weighting          : term frequency (tf)
```

Nga përshkrimi i matricës për korpusin romane vërejmë se 74% e termave të korpusit janë të rralla pra me një frekuence të ulet.

Për të ndërtuar shpërndarjen e fjalëve, fillimisht llogariten frekuencat e fjalëve me ndihmën e funksionit colSums(). Më pas për secilin korpus do të llogariten frekuencat e fjalëve përkatësisht, si me poshtë:

```
> freksh <- colSums(as.matrix(dtmsh))
> length(freksh)
[1] 43928
> rendit <- order(freksh)
> freksh[tail(rendit)]
ishte   nga   nuk   se   për   që   dhe   një   i   në   të
3301    4994  5264 5580 6181   8555   8709   9219  10196 13984 33941
```

Vërehet se për korpusin “shqip” në të cilin janë përfshirë të gjithë tekstet shqip me gjithsej 43928 fjalë të ndryshme, njëmbëdhjetë fjalët me të përdorura janë fjalët: “të”, “në”, “i”, “një”, “që”, “dhe”, “për”, “se”, “nuk”, “nga”, “ishte” të cilat përgjithësisht janë nyje ose lidhëza të gjuhës shqipe. Këto fjalë janë rreth 25.65% e gjithë fjalëve në korpus. Përgjithësisht ato janë fjalë me dy dhe me tre germa. Për ti modeluar këto fjalë si

zinxhir Markovi duhet të llogariten probabilitete kalimi me një dhe me dy hapa, për shembull figura 1.9.

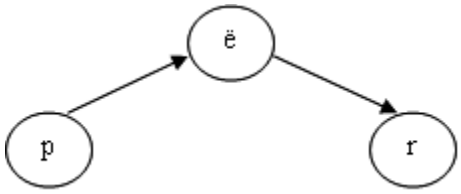

FIGURA 2. 1: FJALA PËR SI ZINXHIR MARKOVI

Kjo shpërndarje e fjalëve më të shpeshta shpjegon rezultatet e kapitullit të parë që modeli më i mirë i zinxhirëve të Markovit ishte Zinxhiri i Markovit i rendit të dytë.

Ndër këto fjalë lidhëzat "që" dhe "se" janë lidhëza të rëndësishme të cilat konsiderohen si lidhëza shumëfunksionale, një klasë e rëndësishme e lidhjes sintaksore gjuhësore të rëndësishme të shqipes, Rushiti(2012). Këto dy lidhëza ndeshen shpesh si lidhëza ftilluese, shkakore, krahasore, qëllimore, kohore, rrjedhimore, etj. Prania e këtyre dy lidhëzave ftilluese në gjuhën shqipe jep mundësi për variacione sintaksore dhe stilistike në ndërtimin e periudhave me fjali të varura. Lidhëza është pjesë e pandryshueshme e ligjëratës, që lidh dy gjymtyrë fjalie ose dy fjali dhe shpreh marrëdhënie të ndryshme sintaksore ndërmjet tyre. Lidhëzat *që* dhe *se* kanë rëndësi të madhe si element i domosdoshëm i komunikimit për të shprehur pse-në e gjërave dhe janë nga shkakoret më të përdorshme. Ato përdoren në gjuhën e folur dhe në atë të shkruar. Lidhëza *që* haset më shumë tek shkrimtarët që shkruajnë në dialektin Tosk ndërsa lidhëza *se* haset më shumë në shkrimtarët që shkruajnë në dialektin Gegë. Për sa përshkruam më sipër, këtyre dy lidhëzave duhet tu kushtohet vëmendje e veçante për identifikimin e autorësisë së teksteve duke i përfshirë ato si elemente kryesore në modelime matematikore.

Për të pasur më të qartë shpërndarjen e fjale të tjera në gjuhen shqipe nëpërmjet funksionit tm_map() dhe removeWords() i përjashtohen nyjet nga korpusi.

Më poshtë po paraqesim fjalët më të përdorura përkatesisht në dy korpuset "artikuj" dhe "romane", të cilat rezultojnë të njëjtat fjalë si në korpusin "shqip".

```
> freq <- colSums(as.matrix(dtm))
> length(freq)
[1] 19246
> rend <- order(freq)
> freq[tail(rend)]
është   nga    nuk   një   dhe   për
 1101  1181  1390  2117  2259  2278
```

```
> freqr <- colSums(as.matrix(dtmr))
> length(freq)
[1] 19299
> renditje <- order(freqr)
> freqr[tail(renditje)]
kishte    nga    nuk    për    dhe    një
  3031   3813   3874   3905   6451   7102
```

Me poshtë me ndihmen e funksionit findFreqTerms() nxjerrim fjalet me frekuence të pakten një vlere të caktuar. Kemi paraqitur fjalet me frekuence të pakten 1000, 500, 200 dhe 100. Fjalet paraqiten në rend alfabetik.

```
> findFreqTerms(dtmsh, lowfreq=1000)
 [1] "ajo"    "apo"    "ata"    "atë"    "dhe"    "duke"   "edhe"   "është"  "gjithë" "is
hte"  "këtë"
[12] "kishte" "kjo"    "kur"    "mos"    "mund"   "nga"    "një"    "nuk"    "pas"    "pë
r"    "për"
[23] "por"    "prej"   "saj"    "shumë"  "tha"    "tij"    "tyre"   "unë"    "vetëm"
> findFreqTerms(dtm, lowfreq=200)
 [1] "ajo"     "apo"     "ata"     "atë"     "bërë"    "dhe"     "duhet"   "duke"    "edhe"
"është"
[11] "gjithë"  "ishte"   "janë"    "kanë"    "këtë"    "këto"    "kishte"  "kjo"     "kur"
"mund"
[21] "nëse"    "nga"     "një"     "nuk"     "parë"    "pas"     "për"     "por"     "prej"
"shumë"
[31] "t'i"     "tij"     "tjetër" "tyre"    "vetëm"

> findFreqTerms(dtmr, lowfreq=800)
 [1] "ajo"    "ata"    "atë"    "dhe"    "duke"   "edhe"   "është"  "gjithë" "ishin"  "is
hte"  "këtë"
[12] "kishte" "kjo"    "kur"    "mos"    "mund"   "nga"    "një"    "nuk"    "pas"    "pë
r"    "por"
[23] "prej"   "saj"    "tha"    "tij"    "tyre"   "unë"    "vetëm"
```

Në të tre rezultatet e mësipërme vërejmë që në çfarëdolloj korpusi të çfarëdolloj madhësie janë thuajse të njëjtat fjalë që kanë frekuencën me të lartë. Kështu që më pas do shqyrtohet vetëm korpusi "shqip"

Një paraqitje tjetër e frekuencave të fjalëve në korpus bëhet nëpërmjet paraqitjes grafike me ndihmën e funksioneve ggplot(), geom_bar() dhe theme() si më poshtë. Grafiku paraqitet në figurën 2.1.

```
> p1<-ggplot(subset(wfreq,freksh>1500),aes(word,freksh))
> p1<-p1+geom_bar(stat="identity")
> p1<-p1+theme(axis.text.x=element_text(angle=45,hjust=1))
> p1
```

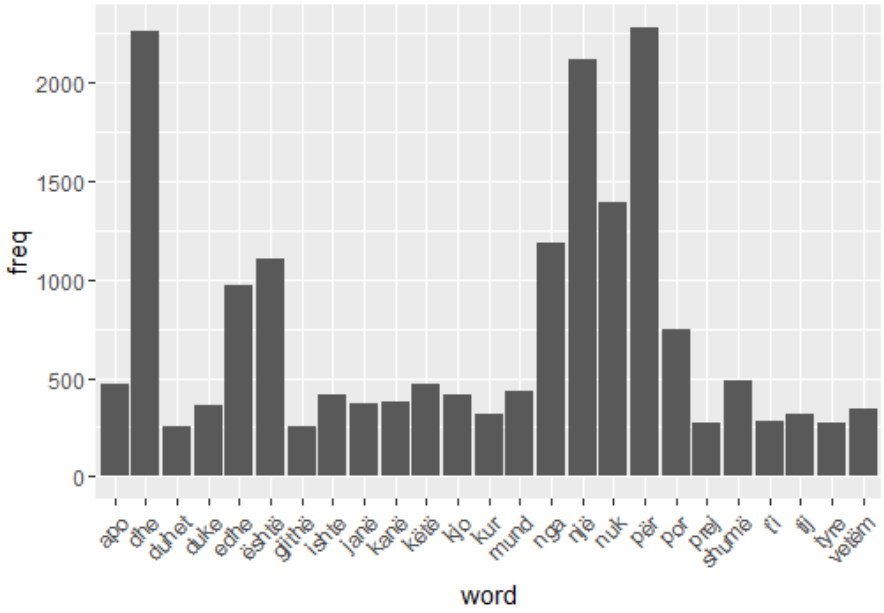

FIGURA 2. 2: GRAFIKU I SHPERNDARJES SË FJALËVE MË TË SHPESHTA SHQIP

Një mënyrë tjetër qe mundëson R për paraqitjen e fjalëve me frekuence me të larte është paraqitja grafike Reja e fjalëve e cila realizohet nëpërmjet funksionit wordcloud(). në figurën 2.2 paraqitet reja e fjalëve me frekuence të paktën 500 për korpusin *shqip*.

```
> library(wordcloud)
> set.seed(142)
> wordcloud(names(freksh), freksh, min.freq=500)
```

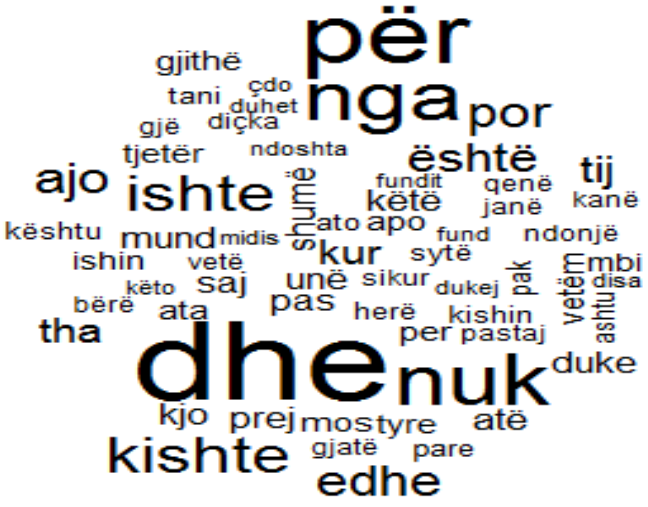

FIGURA 2. 3: REJA E FJALËVE NË KORPUSIN SHQIP

Në figurën 2.2 paraqitet reja e fjalëve me frekuencë të paktën 500 dhe në figurën 2.3 paraqitet reja e fjalëve me frekuencë të paktën 700 por duke shtuar edhe ngjyra në grafik, për të patur më të qarte nga ana vizive fjalët me frekuencë më të lartë.

```
> set.seed(142)
> wordcloud(names(freksh), freksh, min.freq=700, scale=c(5, .1), colors=brewer.pal(6, "Dark2"))
```

FIGURA 2. 4: REJA E FJALËVE NË KORPUSIN SHQIP ME FREKUENCË TË PAKTËN 700

Fjalët sipas frekuencave mund të grupohen në klasa bazuar në algoritmet hierarkikë të analizës së grupimeve, duke i paraqitur grafikisht me ndihmën e dendogramës. Fillimisht pasi largojmë 15% të fjalëve me denduri më të vogël nëpërmjet funksionit hclust() ndërtohet dendograma e cila i grupon fjalet ne 5 klasa si ne figuren 2.4.

```
> library(tm)
> dtmshs <- removeSparseTerms(dtmsh, 0.15)
> dtmshs
<<DocumentTermMatrix (documents: 107, terms: 24)>>
Non-/sparse entries: 2396/172
Sparsity           : 7%
Maximal term length: 6
Weighting          : term frequency (tf)
> library(cluster)
> ds <- dist(t(dtmshs), method="euclidian")
> fit <- hclust(d=ds, method="ward.D")
> fit

Call:
hclust(d = d, method = "ward.D")
```

```
Cluster method   : ward.D
Distance         : euclidean
Number of objects: 18
> plot.new()
> plot(fits, hang=-1)
> groups <- cutree(fit, k=5)
> rect.hclust(fits, k=5, border="red")
```

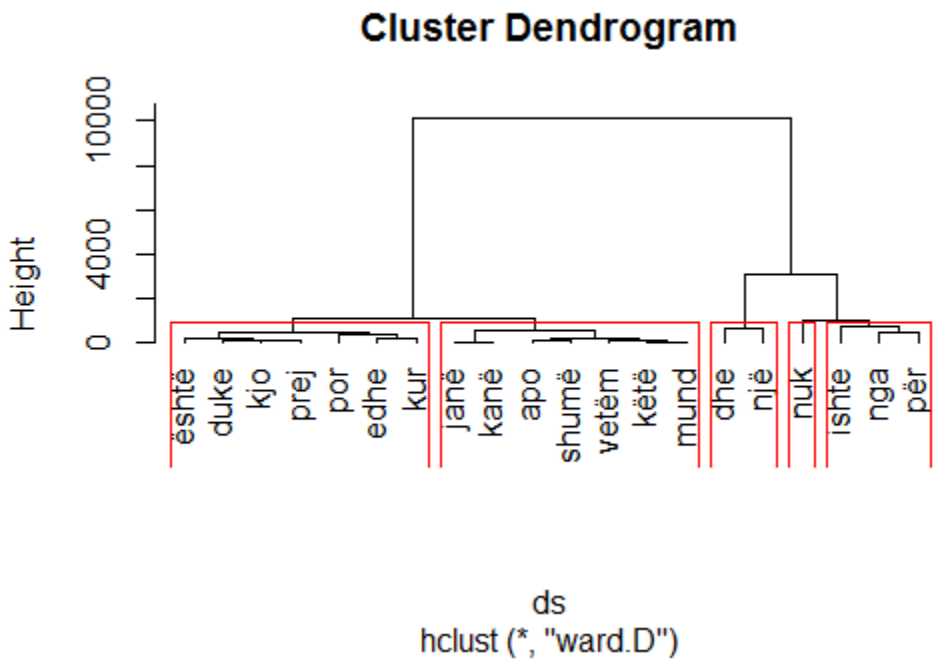

FIGURA 2. 5: DENDOGRAMA E FJALEVE NE KORPUSIN SHQIP

Në këtë mënyrë kemi bërë një ndarje të fjalëve shqip në grupe sipas ndryshimit në frekuenca me njëra tjetrën. Vërehen dy grupe kryesore të fjalëve të cilët ndryshojnë shumë me njëri tjetrin. I pari i cili përbehet nga tre grupime të tjerë qe përmbajnë fjalët me denduri më të madhe dhe i dyti i cili ndahet në dy grupe të tjerë. Pra fjalët i kemi ndare në 5 grupe gjithsej për nga ndryshimi. Fjala "*nuk*" ndryshon më shumë nga grupet e tjera. Ajo ndryshon më pak me grupin e fjalëve *ishte*, *nga*, *për*, dhe pak më shumë ndryshon nga fjalët *dhe* , *një* , të cilat formojnë një grup që ndryshon më shumë nga të gjithë grupet e tjerë.

Meqenëse u ndërtua shpërndarjen e fjalëve sipas frekuencave bëhet e mundur llogaritja e gjatësisë mesatare të fjalëve në korpusin shqip si më poshtë:

```
> freksh <- colSums(as.matrix(dtmsh))
> freksh <- sort(colSums(as.matrix(dtmsh)), decreasing=TRUE)
> WFrek <- data.frame(Fjala=names(freksh), freq= freksh)
> View(WFrek)
> sum(WFrek$freq)
[1] 428476
> library(stringr)
> NRCH <- str_count(WFrek$Fjala)
> GJFJ <- data.frame(NRCH,WFrek$freq)
> View(GJFJ)
> SHUMA <- sum(GJFJ$NRCH*GJFJ$WFrek.freq)
> SHUMA
[1] 2209154
> GJMES <- SHUMA/sum(WFrek$freq)
[1] 5.15584
```

Pra duke u bazuar në frekuencat e fjalëve në korpusin *shqip* vlerësohet gjatësia mesatare e fjalëve në gjuhën e shkruar shqip rreth 5 me mesatare aritmetike 5.16 karaktere.

2.6 Shpërndarja e germave dhe e fjalëve në tekstet shqip

Shpërndarja e fjalëve në tekstet të gjuhëve të ndryshme i bindet ligjit zipf. Në këtë paragraf do të shohim nëse edhe fjalët në gjuhën e shkruar shqipe i binden ligjit Zipf.

Për të krijuar një ide lidhur me shpërndarjen e fjalëve në tekstet e gjuhës shqipe, ndërtojmë shpërndarjen Zipf si më poshtë:

```
> Zipf_plot(dtmsh)
(Intercept)           x
  11.366430   -1.089988
```

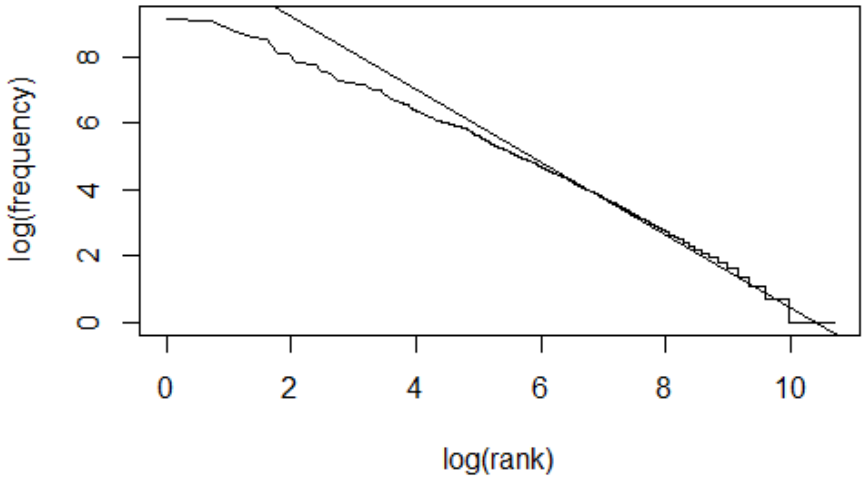

FIGURA 2. 6: LIGJI ZIPF PER FJALET NE KORPUSIN SHQIP

Në figurën 2.5 tregohet paraqitja grafike për shpërndarjen Zipf të fjalëve në korpusin shqip. Drejtëza në grafik është drejtëza sipas ligjit Zipf, vija e thyer paraqet grafikun për fjalët. Vërehet një shmangie nga ligji Zipf për fjalët me frekuence të lartë.

Për korpusin "artikuj" grafiku për ligjin zipf jepet në figurën 2.6.

```
> zipf_plot(dtm)
(Intercept)           x
  8.5258877  -0.8941336
```

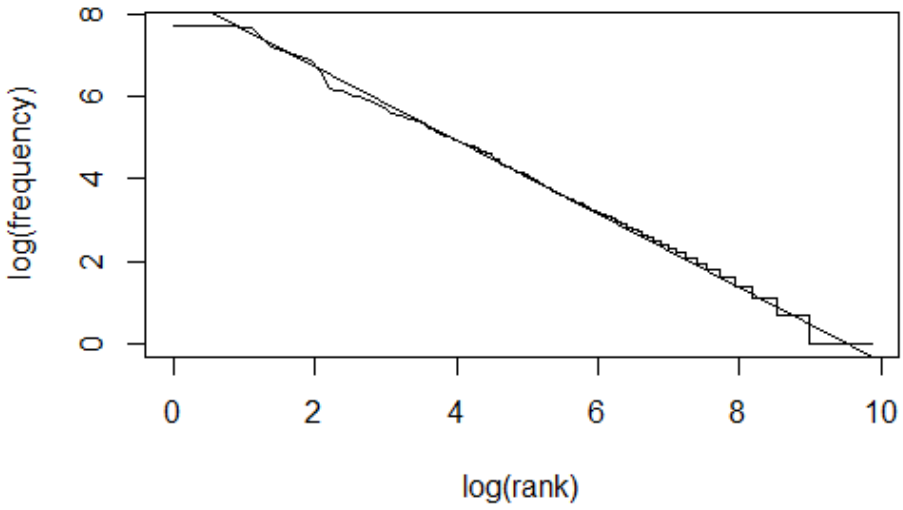

FIGURA 2. 7: LIGJI ZIPF PER FJALET NE KORPUSIN ARTIKUJ

Nga figura 2.6 vërehet se shpërndarja e fjalëve ne korpusin artikuj i përafrohet shumë mirë ligjit Zipf.

Më poshtë do të ndërtoj një shpërndarje për germat në tekstet shqip.

```
> library(stringr)
> freqgerma<-c(str_count(shqip,"a"),str_count(shqip,"b"),str_count(shqip,"c"),str_count(shqip,"ç"),str_count(shqip,"d"),str_count(shqip,"dh"),str_count(shqip,"e"),str_count(shqip,"ë"),str_count(shqip,"f"),str_count(shqip,"g"),str_count(shqip,"gj"),str_count(shqip,"h"),str_count(shqip,"i"),str_count(shqip,"j"),str_count(shqip,"k"),str_count(shqip,"l"),str_count(shqip,"ll"),str_count(shqip,"m"),str_count(shqip,"n"),str_count(shqip,"nj"),str_count(shqip,"o"),str_count(shqip,"p"),str_count(shqip,"q"),str_count(shqip,"r"),str_count(shqip,"rr"),str_count(shqip,"s"),str_count(shqip,"sh"),str_count(shqip,"t"),str_count(shqip,"th"),str_count(shqip,"u"),str_count(shqip,"v"),str_count(shqip,"x"),str_count(shqip,"xh"),str_count(shqip,"y"),str_count(shqip,"z"),str_count(shqip,"zh"))

>alfabeti<-
c("a","b","c","ç","d","dh","e","ë","f","g","gj","h","i","j","k","l","ll","m","n","nj","o","p","q","r","rr","s","sh","t","th","u","v","x","xh","y","z","zh")
```

```
>freqshkronja<-c(176454, 26254, 9320, 7320, 80957, 19636, 225763, 223256, 20625, 36931,
15102, 114609, 180649, 91043, 81410, 52875, 10395, 84483, 151560, 17245, 85527, 68320,
24892, 164782, 10279, 139098, 57840, 208123, 11893, 89207, 30718, 2060, 1012, 17956,
16847, 1917)
>freqalfabeti<-data.frame(alfabeti,freqshkronja)
>o=order(freqshkronja)
>a=freqalfabeti[o,]
>a
```

Alfabeti	Freqshkronja	alfabeti	freqshkronja
E	225763	G	36931
Ë	223256	V	30718
T	208123	B	26254
I	180649	Q	24892
A	176454	F	20625
R	164782	Dh	19636
N	151560	Y	17956
S	139098	Nj	17245
H	114609	Z	16847
J	91043	Gj	15102
U	89207	Th	11893
O	85527	Ll	10395
M	84483	Rr	10279
K	81410	C	9320
D	80957	Ç	7320
P	68320	X	2060
Sh	57840	Zh	1917
L	52875	Xh	1012

TABELA 2. 1: SHPERNDARJA E SHKRONJAVE NE KORPSIN SHQIP

Nga tabela vërehet se germat me të përdorura në korpusin tonë janë e, ë dhe t të cilat përbejnë 25.7% të gjithë germave në korpus duke ju shtuar edhe germat i, a, r, n së bashku përbëjnë 52% të germave të korpusit. Gjithashtu vërejmë se 39% e germave në korpus janë zanore.

```
> library(stringr)
>freqgerma<-
c(str_count(dok,"a"),str_count(dok,"b"),str_count(dok,"c"),str_count(dok,"ç"),str_count(dok,"d"),str_count(dok,"dh"),str_count(dok,"e"),str_count(dok,"ë"),str_count(dok,"f"),str_count(dok,"g"),str_count(dok,"gj"),str_count(dok,"h"),str_count(dok,"i"),str_count(dok,"j"),str_count(dok,"k"),str_count(dok,"l"),str_count(dok,"ll"),str_count(dok,"m"),str_count(dok,"n"),str_count(dok,"nj"),str_count(dok,"o"),str_count(dok,"p"),str_count(dok,"q"),str_count(dok,"r"),str_count(dok,"rr"),str_count(dok,"s"),str_count(dok,"sh"),str_count(dok,"t"),str_count(dok,"th"),str_count(dok,"u"),str_count(dok,"v"),str_count(dok,"x"),str_count(dok,"xh"),str_count(dok,"y"),str_count(dok,"z"),str_count(dok,"zh"))
```

```
>freqshkronjaA<-
c(42805,6305,2585,1355,19394,5044,56434,57120,4878,7678,3511,24888,46472,20368,21720,1374
0,2203,21100,37560,3688,23524,18836,6799,41320,1900,34353,13017,52350,1982,20039,8134,419
,172,3573,4221,413)
> freqalfabetiA<-data.frame(alfabeti,freqshkronjaA)
> o=order(freqshkronjaA)
> AAa=freqalfabetiA[o,]
> AAa
```

Alfabeti	freqshkronjaA	Alfabeti	freqshkronjaA
Ë	57120	V	8134
E	56434	G	7678
T	52350	Q	6799
I	46472	B	6305
A	42805	Dh	5044
R	41320	F	4878
N	37560	Z	4221
S	34353	Nj	3688
H	24888	Y	3573
O	23524	Gj	3511
K	21720	C	2585
M	21100	Ll	2203
J	20368	Th	1982
U	20039	Rr	1900
D	19394	Ç	1355
P	18836	X	419
L	13740	Zh	413
Sh	13017	Xh	172

TABELA 2. 2: SHPERNDARJA E SHKRONJAVE NE KORPUSIN ARTIKUJ

Vërehet se shpërndarja e germave në tekstet artikuj është thuajse njësoj si më sipër, pra rreth 26% të germave e përbëjnë germat *ë, e* dhe *t*, 7 germat *e, ë, t,i,a,r,n* përbëjnë rreth 53% të germave dhe 39.6% e germave janë zanore. Vërehet e njëjta shpërndarje e germave si edhe në korpusin shqip. Atëherë si konkluzion numrin e zanoreve, numrin e germave me denduri më të larte mund ti përdorim si elemente të rëndësishëm për identifikimin e autorësisë së teksteve shqip.

Rezultate mbi shpërndarjen e germave në tekste shqip janë diskutuar edhe nga Dika(2012) i cili ka paraqitur përafërsisht të njëjtat rezultate pavarësisht se korpusi i tyre është një korpus tekstesh të çfarëdoshme shqip. Pra në përgjithësi janë të njëjtat germa dhe fjalë që kanë frekuencë më të madhe por ndryshojnë renditje nga teksti në tekst.

Meqenëse u ndërtuan shpërndarja e germave në dy korpuset *shqip* dhe *artikuj* të shohim nëse shpërndarja e germave i bindet ligjit Zipf.

Duke shënuar ZArt të dhënat mbi renditjen e frekuencave të germave në korpusin për artikujt dhe Ztekste të dhënat mbi renditjen e frekuencave të germave në korpusin shqip ndërtoj shpërndarjen Zipf si më poshtë:

```
> Zipf_plot(ZArt)
(Intercept)           x
  13.353316   -9.885385
```

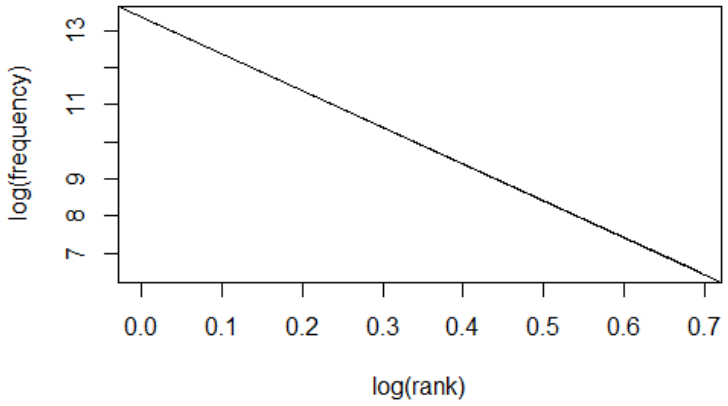

FIGURA 2. 8: LIGJI ZIPF PËR GERMAT NË KORPUSIN ARTIKUJ

```
> Zipf_plot(Ztekste)
(Intercept)           x
   14.75409   -11.90628
```

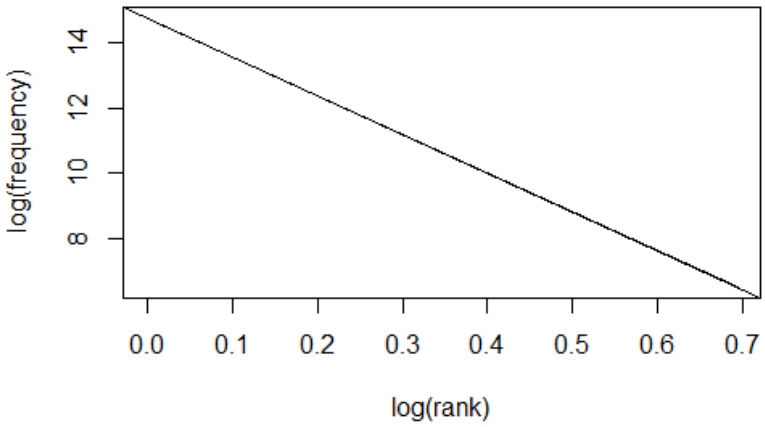

FIGURA 2. 9 : LIGJI ZIPF PËR GERMAT NË KORPUSIN SHQIP

Nga figurat 2.7 dhe 2.8 vërehet se shkronjat shqip i binden ligjit Zipf.

2.7 Variabla sasiore nga tekstet shqip

Meqenëse në paragrafin e mësipërm pasi u tregua një shpërndarje të germave në tekstet shqip u përcaktuan numri i zanoreve, lidhëzat *që* dhe *se* si elemente të rëndësishëm gjuhësor e stilistikë. Element stili për një autor është edhe gjatësia e fjalive, pra numri i fjalëve ne një fjali për rrjedhoje edhe numri i fjalëve ne tekstet e tij. Pra për identifikimin e autorit lind nevoja e vrojtimit të variablave të ndryshëm nga tekstet shqip me qëllim ndërtimin e modeleve statistikore për identifikimin e autorit.

Për të gjithë tekstet shqip vrojtohen numri i fjalëve, numri i germave, numri i zanoreve, numri i bashkëtingëlloreve, numri i fjalive, numri i shenjave të pikësimit.

Meqenëse tekstet ndryshojnë për nga madhësia ato shqyrtohen në dy grupe, artikuj dhe romane. Shënoj "dataT" dhe "dataR" vrojtimet përkatësisht për artikujt dhe romanet.

```
> paper=read.table(file="clipboard",header=TRUE)
> summary(paper)
Nr_fjale     Nr_germa      Nr_bashketingellore Nr_zanore    Nr_shenjapikesimi   Nr_fjali
 Min.   : 605   Min.   : 2959   Min.   : 1782   Min.   :1177   Min.   : 77.0   Min.   : 10.00
 1st Qu.:1006   1st Qu.: 4808   1st Qu.: 2920   1st Qu.:1916   1st Qu.:120.0   1st Qu.: 41.00
 Median :1152   Median : 5600   Median : 3344   Median :2268   Median :141.5   Median : 51.00
 Mean   :1280   Mean   : 6216   Mean   : 3723   Mean   :2493   Mean   :156.6   Mean   : 54.71
 3rd Qu.:1440   3rd Qu.: 7154   3rd Qu.: 4256   3rd Qu.:2877   3rd Qu.:177.5   3rd Qu.: 62.75
 Max.   :3505   Max.   :17836   Max.   :10535   Max.   :7301   Max.   :364.0   Max.   :146.00

> dataRomane=read.table(file="clipboard",header=TRUE)
> summary(dataRomane)
    Nr_fjale         Nr_germa        Nr_zanore      Nr_bashketingellore    Nr_fjali
 Min.   : 16642   Min.   : 74098   Min.   : 28836   Min.   : 45262      Min.   : 1119
 1st Qu.: 34102   1st Qu.:152668   1st Qu.: 59881   1st Qu.: 92787      1st Qu.: 2284
 Median : 41147   Median :183702   Median : 71446   Median :112256      Median : 3302
 Mean   : 59924   Mean   :264242   Mean   :106632   Mean   :157610      Mean   : 4236
 3rd Qu.: 67859   3rd Qu.:304166   3rd Qu.:121363   3rd Qu.:182803      3rd Qu.: 5044
 Max.   :157760   Max.   :678228   Max.   :283655   Max.   :394573      Max.   :10577
```

Artikujt e shqyrtuar janë mesatarisht me 1280 fjalë dhe 6216 germa secili ndërsa romanet janë mesatarisht me 59924 fjalë dhe 264242 germa.

Në paragrafin 2.5 u vlerësua duke u bazuar në shpërndarjen e fjalëve gjatësia mesatare e fjalës në gjuhën e shkruar shqip rreth 5 germa. Meqenëse janë vrojtuar numri i fjalëve dhe numri i germave në secilin tekst gjatësia mesatare e fjalës mund të vlerësohet si raport i numrit te germave me numrin e fjalëve në një korpus. Në të dy rastet vlerësojmë

mesatarisht 4-5 germa për fjalë pra mesatarisht gjatësia e fjalëve në gjuhën shqipe është 4-5 germa duke e vlerësuar atë si raport të numrit të germave me numrin e fjalëve në tekste. Për artikujt përftohet një interval besimi 95% për numrin mesatar të germave në një fjale] 4.803766; 4.894367[.

```
> View(paper)
> gf=paper$N_letters/paper$N_words
> t.test(gf,mu=4.85)

         One Sample t-test

data:  gf
t = -0.040904, df = 99, p-value = 0.9675
alternative hypothesis: true mean is not equal to 4.85
95 përcent confidence interval:
 4.803766 4.894367
sample estimates:
mean of x
 4.849066
```

Për të gjithë tekste shqip vlerësohet intervali besimi 95%,]4.776375; 4.869408[për numrin mesatar të germave në një fjale:

```
> teksteshqip=read.table(file="clipboard",header=TRUE)
> g_f<-teksteshqip$NrGerma/teksteshqip$NrFjale
> summary(g_f)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  4.299   4.640   4.841   4.823   4.998   5.359
> t.test(g_f,mu=4.8)

         One Sample t-test

data:  g_f
t = 0.97567, df = 106, p-value = 0.3315
alternative hypothesis: true mean is not equal to 4.8
95 percent confidence interval:
 4.776375 4.869408
sample estimates:
mean of x
 4.822891
```

Nga të dhënat vërehet se në të dy llojet e teksteve si artikuj dhe romane ruhet raporti i zanoreve dhe bashkëtingëlloreve në tekste. Në të gjithë tekstet rreth 40% e germave janë zanore dhe 60% janë bashkëtingëllore. Më poshtë behet vlerësim pikësor dhe intervalor me besueshmëri 95% për raportin e zanoreve me germat ne tekstet shqip.

```
> z_g=teksteshqip$NrZanore/teksteshqip$NrGerma
> summary(z_g)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.3236  0.3975  0.4027  0.4006  0.4069  0.4182
> t.test(z_g,mu=0.4)

         One Sample t-test

data:  z_g
t = 0.5018, df = 106, p-value = 0.6168
alternative hypothesis: true mean is not equal to 0.4
95 përcent confidence interval:
 0.3983096 0.4028361
sample estimates:
mean of x
0.4005728
```

Në gjuhën angleze rreth 37% e germave ne tekste janë zanore Dika(2012).

Një koeficient i cili është shumë i rëndësishëm për studiuesit e fonetikës është raporti i bashkëtingëlloreve me zanoret për një gjuhë të caktuar si një tregues i thjeshtësisë së të kuptuarit dhe të folurit të saj. Raporti i bashkëtingëlloreve dhe i zanoreve në një gjuhë është një element i rëndësishëm për shqiptimin e fjalës Orr et al. (2010) dhe Port&Dalby (1982). Sa më afer 1 të jetë ky koeficient aq më lehtë shqiptohet një gjuhë. Ky koeficient gjithashtu përdoret gjerësisht në ndërtimin e aparateve të dëgjimit Kennedy et al. (1998). Meqenëse u kostatua një raport konstant midis zanoreve dhe bashkëtingelloreve në gjuhën e shkruar shqipe, më poshtë vlerësohet raporti i numrit të bashketingelloreve me numrin e zanoreve në tekstet shqip.

```
> b_z=teksteshqip$NrBashketingellore/teksteshqip$NrZanore
> summary(b_z)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.391   1.458   1.483   1.499   1.515   2.091
> t.test(b_z,mu=1.5)

         One Sample t-test

data:  b_z
t = -0.13638, df = 106, p-value = 0.8918
alternative hypothesis: true mean is not equal to 1.5
95 përcent confidence interval:
 1.482676 1.515094
sample estimates:
mean of x
 1.498885
```

Si rezultat ky koeficient për gjuhen e shkruar shqipe vlerësohet me mesatare 1.499 dhe me një interval besimi 95% ,] 1.482676 1.515094 [. Për gjuhen angleze ky koeficient është afërsisht 1.72 dhe për gjuhen Portugeze është afërsisht 1.14 Pollo et al. (2005). Sipas këtyre rezultateve gjuha Shqipe rezulton më e lehtë për tu shqiptuar se gjuha Angleze dhe më e vështirë se gjuha Portugeze. Për gjuhen Italiane ky koeficient është 1.148, për gjuhen Gjermane është 1.76 dhe për gjuhen Franceze është 1.37.

2.9 Përfundime

Në paragrafët e mësipërm u bë një analize statistikore e germave dhe fjalëve në tekstet shqip. Shpërndarja e germave dhe e fjalëve në tekstet shqip i bindet ligjit Zipf. Fjalët që shfaqen më shpesh në tekstet shqip janë përgjithësisht nyje dhe lidhëza të gjuhës shqipe të cilat përbëjnë rreth 25.65% të fjalëve në korpus. Fjalët me të shpeshta në gjuhën e shkruar shqipe janë kryesisht fjalë me dy dhe me tre germa fakt i cili shpjegon rezultatet e kapitullit të parë që modeli më i mirë i zinxhirëve të Markovit ishte Zinxhiri i Markovit i rendit të dytë. U vlerësua gjatësia mesatare të fjalëve në gjuhën shqipe të shkruar në dy mënyra, ku si rezultat gjatësia mesatare në të dy mënyrat rezulton thuajse e njëjtë rreth 5 germa për fjalë. Rreth 40% e germave janë zanore, duke vlerësuar një interval besimi 95% për raportin e zanoreve ndaj numrit të germave]0.3984701; 0.4031952[. Një koeficient i cili është shumë i rëndësishëm për studiuesit e fonetikes është raporti i bashkëtingëlloreve me zanoret për një gjuhë të caktuar si një tregues i thjeshtësisë se kuptuarit dhe të folurit të një gjuhe. Vërehet që edhe në tekstet shqip ky raport ruhet, për artikujt rezulton me mesatare 1.499 dhe me një interval besimi 95%]1.482676; 1.515094[.

Kapitulli 3

Regresi Logjistik

3.1 Hyrje

Problemi i identifikimit të autorit është përpjekja për të gjetur autorësinë e teksteve me autorësi të panjohur ose autorësi të dyshimte. Për një tekst të zgjedhur në duhet të marrim një vendim statistikor nëse ky tekst i takon apo jo të një autori të caktuar. në teorinë e klasifikimit statistikor gjejmë modele të ndryshme statistikore një prej të cilave është edhe regresi logjistik Hastie et al.(2009). Përpara se të bëjmë modelime matematikore duhet të njihemi me variablat dhe shpërndarjen probabilitare të tyre. Në kapitullin e parë u ndërtuan modele të ndryshme të zinxhirëve të Markovit duke u bazuar në n-gram-at e germave të gjuhës shqipe dhe u vlerësua entropia e teksteve nëpërmjet zinxhirëve të Markovit të rendeve të larta. Për të kuptuar arsyen e modelimit me te mire me zinxhir Markovi te rendit te dyte ne kapitullin 2 u bë përpunim, analiza statistikore dhe vrojtim i variablave sasiore në tekstet shqip. Me ndihmën e variablave sasiorë të vrojtuar në kapitullin 2, do të ndërtohen modele të regresit logjistik me qellim identifikimin e autorit të teksteve. Në paragrafët pasues do trajtohet teoria e regresit logjistik dhe llojet e tij sipas Hosmer, Lemeshow (2013), përshtatja e modeleve për identifikimin e autorit te teksteve, do të ndërtohen dhe diskutohen modele të ndryshme të regresit logjistik të shumëfishtë dhe multinomial për identifikimin e autorit në tekstet shqip dhe së fundi do të diskutohen avantazhet dhe disavantazhet e secilit model.

3.2 Regresi i thjeshtë logjistik

Modelet e regresit janë bërë një komponentë integrale e analizës së të dhënave të cilat duan të përshkruajnë lidhjen midis një variabli të varur dhe disa variablave të pavarur. Ndodh shpesh që variabli i varur të jete diskret i cili merr vetëm me dy ose më shumë vlera të mundshme. Gjatë dekadës së kaluar në shumë fusha studimi, modeli i regresit logjistik është kthyer në metodën bazë për analizën e të dhënave në këtë situatë. Përpara se të bëjmë një studim mbi regresin logjistik është shumë e rëndësishme të kuptojmë që qëllimi i një analize duke përdorur këtë metodë është e njëjtë me teknikat e tjera të modelimit statistikor: të gjejmë modelin me të përshtatshëm statistikisht i cili shpjegon lidhjen midis një variabili të varur dhe një bashkësie me variabla të pavarur ose shpjegues. Ajo ç'ka e dallon modelin e regresit logjistik nga modelet e tjera statistikore, si modeli i regresit linear është se variabli i varur është cilësor (kategorik). Ky dallim ndërmjet regresit logjistik dhe atij linear reflektohet në zgjedhjen e një modeli parametrik dhe në supozimet e modelit. Teknikat e përdorura në analizën e regresit linear motivojnë përqasjet për ndertimin e modelit të regresit logjistik sipas Hosmer, Lemeshow (2013) . Në çdo problem të regresit, vlerësues sasior kryesor është vlera mesatare e variablit të varur, kur njihet vlera e variablit të pavarur. Kjo madhesi quhet ***pritje matematike***

(mesatare) e kushtezuar e cila shënohet " $E(Y / x)$" ku Y tregon variablin e varur dhe x tregon një vlerë të varablit të pavarur, por në regresin logjistik mund të perdoret ky shenim vetëm në qoftë se njohim ligjin probabilitar të Y.

Për analizën e variablit të varur kategorik janë propozuar shumë funksione shpërndarje. Shpërndarja probabilitare që përdoret për variabla kategorik është shpërndarja logjistike. Janë dy arsye kryesore për zgjedhjen e shpërndarjes logjistike. E para, nga pikëmaja matematikore, është një funksion fleksibël dhe lehtësisht i përdorshëm, dhe së dyti, e lejon veten drejt një interpretimi në kuptimin rastësor.

Shënojmë $\pi(x) = E(Y/x)$ kur Y ka shpërndarje logjistike. Modeli i regresit logjistik merr trajtën:

$$\pi(x) = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}} \qquad (3.1)$$

Një trasformim i $\pi(x)$ që është shumë thelbësor në studimin e regresit logjistik është trasformimi logit. Ky trasformim përcaktohet, në termat e $\pi(x)$ si:

$$g(x) = \ln\left[\frac{\pi(x)}{1-\pi(x)}\right] = \beta_0 + \beta_1 x.$$

Rëndësia e këtij transformimi është se $g(x)$ shqyrtohet si një modeli i regresit linear. Logit, g(x), është funksion linear i parametrave, mund të jetë i vazhdueshëm, dhe me vlera reale të varura nga madhësia e x.

Vlera e një variabli të varur kategorik me dy vlera mund të shprehet si $y = \pi(x) + \varepsilon$ ku madhësia ε mund të supozohet si një nga dy vlerat e mundshme. Nëse y=1, atëherë $\varepsilon = 1 - \pi(x)$ me probabilitet $\pi(x)$, dhe nëse $y = 0$ atëherë $\varepsilon = -\pi(x)$ me probabilitet $1 - \pi(x)$.

Y	0	1
E	$-\pi(x)$	$1 - \pi(x)$
P	$1 - \pi(x)$	$\pi(x)$

TABELA 3. 1: LIGJI PROBABILITAR I MADHESISE ε

Kështu që madhësia ε ka shpërndarje me mesatare zero dhe variancë të barabartë me $\pi(x)[1 - \pi(x)]$. Kjo do të thotë se shpërndarja e kushtëzuar e një variabli të varur ndjek një shpërndarje binomiale me një probabilitet të dhënë nga mesatarja e kushtëzuar $\pi(x)$.

Pra në analizën e regresit kur variabli i varur është kategorik plotësohen këto kushte :

Mesatarja e kushtëzuar e ekuacionit të regresit përcaktohet me vlera nga 0 deri në 1.

Shpërndarja termave të gabimeve është binomiale e cila do të jetë një shpërndarje statistikore mbi të cilën është bazuar analiza e tyre.

Parimet që drejtojnë një analizë e regresiot linear drejtojne gjithashtu në regresin logjistik.

3.2.1 Përshtatja e Modelit të Regresit Logjistik

Supozojmë se kemi një zgjedhje me *n* vrojtime të pavarura të çiftit $(x_{i,}y_i)$, *i= 1, 2..., n,* ku y_i përfaqëson vlerën e një variabli të varur kategorik dhe $x_{i,}$ është vlera e i-te e variablit të pavarur. Për më tepër, supozojmë se variabli i varur merr vetëm dy vlera 0 ose 1, duke përfaqësuar mungesën ose prezencën e një karakteristike. Për të përshtatur modelin e regresit logjistik në ekuacionin 3.1 për një bashkësi të dhënash kërkohet të përcaktohen vlerat β_0 dhe β_1 të parametrave të panjohura.

Në regresin linear për të vlerësuar parametrat e panjohura përdoret metoda e katrorëve më të vegjël dhe minimizimi i shumës se katroreve nëpërmjet funksionit të përgjasisë maksimale. Gjithashtu për vlerësimin e parametrave të regresit logjistik përdoret funksioni i përgjasisë maksimale. Metoda e përgjasisë maksimale përfton vlera të parametrave të panjohur të cilat maksimizojnë probabilitetin qe të kemi të dhënave të vrojtuara. Me qëllim aplikimin e metodës së përgjasisë maksimale fillimisht ndërtohet funksioni i përgjasisë. Ky funksion shpreh probabilitetin e të dhënave të vrojtuara si një funksion të parametrave të panjohur. Vlerësuesit e përgjasise maksimale të këtyre parametrave zgjidhen të jenë ato vlera që maksimizojnë këtë funksion. Kështu që, vlesesuesit e parametrave janë më të përafërta me të dhënat e vrojtuara. Më poshtë do të përshkruajmë se si gjenden vlerësuesit e parametrave për modelin e regresit logjistik.

Nëse *Y* merr vetëm dy vlera 0 ose 1 atëherë shprehja për $\pi(x)$ e dhënë në ekuacionin 3.1 siguron (për vektorin e parametrave $\beta = (\beta_0, \beta_1)$, probabilitetin e kushtëzuar që *Y* të marrë vlerën 1, për *x* e dhënë e cila shënohet si $P(Y = 1/x)$. Rrjedhimisht madhësia $1 - \pi(x)$ jep probabilitetin e kushtëzuar që *Y* të marr vlerën 0, për *x* e dhënë, $P(Y = 0/x)$. Kështu që, për çiftin $(x_{i,}y_i)$, kur y_i= 1, për funksionin e përgjasisë kemi $\pi(x_i)$, ndërsa kur $y_i = 0$, për funksionin e përgjasisë kemi $1 - \pi(x_i)$, ku madhesia $\pi(x_i)$, tregon vlerën e $\pi(x)$ të llogaritur në x_i. Një mënyre e përshtatshme për të shprehur kontributin për funksionin të përgjasisë të çiftin$(x_{i,}y_i)$, është nëpërmjet ekuacionit:

$$\pi(x_i)^{y_i}[1 - \pi(x_i)]^{1-y_i} \qquad (3.2)$$

Meqënëse vrojtimet janë të pavarura, funksioni i përgjasisë merret si një produkt në terma të dhëna nga ekuacioni 3.2 si vijon:

$$l(\beta) = \prod_{i=1}^{n} \pi(x_i)^{y_i}[1 - \pi(x_i)]^{1-y_i} \qquad (3.3)$$

Parimi i përgjasisë maksimale përcakton si vlerësues të parametrit β vlerën që maksimizon shprehjen në ekuacionin 3.3. Megjithatë, matematikisht është më e lehtë përdorimi i logaritmit të ekuacionit 3.3 si më poshtë:

$$L(\beta) = \ln[\, l(\beta)] = \sum_{i=1}^{n}\{y_i \ln[\pi(x_i)] + (1 - y_i) \ln [1 - \pi(x_i)]\} \qquad (3.4)$$

Për të gjetur vlerën e β që maksimizon $L(\beta)$ llogariten derivatet e pjesshme të $L(\beta)$ respektivisht lidhur me $\beta_0, dhe\ \beta_1$ dhe barazohen ato me zero. Këto ekuacione të njohura si ekuacione të përgjasise janë:

$$\sum_{i=1}^{n}[y_i - \pi(x_i)] = 0 \qquad (3.5)$$

dhe

$$\sum_{i=1}^{n} x_i[y_i - \pi(x_i)] = 0 \qquad (3.6)$$

Për regresin logjistik shprehjet në ekuacionet 3.5 dhe 3.6 janë jolineare lidhur me $\beta_0, dhe\ \beta_1$, kështu kërkojnë metoda speciale për zgjidhjen e tyre. Këto metoda janë përsëritëse në natyrë dhe janë programuar në software të posaçme të regresit logjistik. Vlera e β që merren si zgjidhje të ekuacioneve (3.5) dhe (3.6) quhet vlerësues i përgjasisë maksimale dhe shënohet $\hat{\beta}$. Përgjithësisht do të përdoret simboli "^" për vlerësuesit e përgjasisë maksimale të një madhësie të caktar. Shënojmë $\hat{\pi}(x_i)$ vlerësuesin e përgjasisë makimale të $\pi(x_i)$. Madhësia $\hat{\pi}(x_i)$ jep një vlerësues të probabilitetit të kushtëzuar që Y të jetë i barabartë me 1, për x e barabartë me x_i. Si e tillë ajo përfaqëson vlerën e parashikuar të modelit të regresit logjistik. Një paraqitje tjeter e ekuacionit (3.5) është:

$$\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \hat{\pi}(x_i)$$

3.2.2 Kontrolli i Rëndësise se Koefiçentëve dhe intervalet e besimit

Pasi vlerësohen koefiçentët në modelin e regresit logjistik është i rëndësishëm kontrolli i rendesise se variablave. Kjo zakonisht përfshin formulimin dhe kontrollin e hipotezave statistikore për të përcaktuar nëse variablat e pavarura në model janë në menyre të rendesishme statistikisht të lidhura me variablin e varur.

Njësoj si në kontrollin e rëndësisë së parametrave në modelin e regresit linear edhe në regresin logjistik përdoret i njëjti parim, ai i krahasimit të vlerave të vrojtuara të variablit me vlerat e variablit të përafruara sipas modelit të ndertuar. Në regresin logjistik krahasimi i vlerave të parashikuara dhe atyre të vrojtuara bazohet në logaritmin e funksionit të përgjasise 3.4. Për të kuptuar më mirë këtë krahësim, është e nëvojshme të mendojmë një vlerë të vrojtuar të variablit të varur dhe një vlerë e parashikuar (vlerësuar) të tij të marre si rezultat i një modeli i ngopur (saturated model). Një model i ngopur është një model që përmban aq parametra sa variabla të vrojtuar.

Krahasimi i vlerave të vrojtuara dhe atyre të parashikuara duke përdorur funksionin e përgjasisë jepet në formulën 3.7:

$$D = -2ln\left[\frac{Përgjasia\ per\ modelin\ e\ pershtatur}{Përgjasia\ per\ modelin\ e\ ngopur}\right] \qquad (3.7)$$

Madhësia brenda kllapave në shprehjen e mësipërme quhet raporti i përgjasisë. Kjo statistike njihet ndryshe me emrin raporti i përgjasise. Duke përdorur ekuacionin 3.4 dhe 3.7, ai merr trajten e meposhtëme:

$$D = -2\sum_{i=1}^{n}[y_i \ln\left(\frac{\hat{\pi}_i}{y_i}\right) + (1 - y_i)\ln\left(\frac{1-\hat{\pi}_i}{1-y_i}\right)] \qquad (3.8)$$

ku $\hat{\pi}_i = \hat{\pi}(x_i)$.

Statistika D, në ekuacionin 3.8 quhet deviancë nga disa autor dhe luan një rol kryesor për të vlerësuar cilësinë e të përshtatjes se modelit e quajtur *përputhshmëria e modelit*. Devianca për regresin logjistik luan të njëjtin rol si shuma e katroreve të gabimeve (SKG) në regresin linear. Në fakt, devianca në ekuacionin 3.8, kur zbatohet për regresin linear, është e barabartë me SKG. Për më tepër, kur vlerat e variablit të varur janë 0 ose 1, përgjasia e modelit të ngopur është e barabarte me 1. Nga përkufizimi i një modeli të ngopur rrjedh që $\hat{\pi}_i = y_i$ dhe përgjasia është:

$$l(modeli\ i\ ngopur) = \prod_{i=1}^{n} y_i^{y_i}(1-y_i)^{(1-y_i)} = 1 \qquad (3.9)$$

Nga ekuacioni 3.7 rrjedh se devianca është:

$$D = -2\ln\left(Përgjasia\ per\ modelin\ e\ pershtatur\right) \qquad (3.10)$$

Për të vlerësuar rëndësinë e variablit të pavarur, krahasohen vlerat e *D-se* me dhe pa variablat e pavarur në ekuacion. Ndryshimi i *D* për shkak të përfshirjes së variablave të pavarur në model shenohet:

$$G = D(modeli\ pa\ variablin) - D(modelin\ me\ variablin)$$

Sipas parimit të përgjasisë statistika G merr trajtën:

$$G = -2ln\left[\frac{përgjasia\ pa\ variablin}{përgjasia\ me\ variablin}\right] \qquad (3.11)$$

Për rastin e një variabli të pavarur, tregohet lehtë se vlerësuesi i përgjasisë maksimale për parametrin β_0 është $\ln(n_1/n_0)$ ku $n_1 = \sum y_i$ dhe $n_0 = \sum(1-y_i)$ dhe vlera e parashikuar është konstante, n_1/n. Në këtë rast vlera e G merr trajtën 3.12:

$$G = -2\ln\left[\frac{(\frac{n_1}{n})^{n_1}(\frac{n_0}{n})^{n_0}}{\prod_{i=1}^{n}\hat{\pi}_i^{y_i}(1-\hat{\pi}_i)^{1-y_i}}\right] \qquad (3.12)$$

ose

$$G = 2\{\sum_{i=1}^{n}[y_i\ln\hat{\pi}_i + (1-y_i)ln(1-\hat{\pi}_i)] - [n_1ln(n_1) + n_0ln(n_0) - nln(n)]\} \qquad (3.13)$$

Statistika G ka shpërndarje Hi-katror (χ^2) me 1 shkallë lirie për hipotezën $H_0: \beta_1 = 0$.

Një tjetër statistikë e ngjashme është testi Wald. Supozimet e nevojshme për këtë test janë të njëjta si ato të testit 3.11. Testi Wald merret duke krahasuar vlerësuesin e përgjasise maksimale $\hat{\beta}_1$ të parametrit β_1, me një vlerësues të gabimit standard (SE) të tij. Testi Wald i cili ka shpërndarje normale standarde jepet nga formula 3.14

$$W = \frac{\widehat{\beta}_1}{\widehat{\mathrm{SE}}(\widehat{\beta}_1)} \qquad (3.14)$$

Duke u bazuar në statistikën Wald ndërtohen edhe intervalet e besimit për nivel rëndësie α të dhene, të parametrave të modelit si me poshtë:

$$\hat{\beta}_1 \pm z_{1-\alpha/2}\widehat{\mathrm{SE}}(\hat{\beta}_1) \qquad (3.15)$$

$$\hat{\beta}_0 \pm z_{1-\alpha/2}\widehat{\mathrm{SE}}(\hat{\beta}_0) \qquad (3.16)$$

Logit është forma lineare e modelit të regresit logjistik dhe si e tillë, ka më shumë ngjashmëri me modelin e regresit linear. Vlerësuesi për modelin logit është:

$$\hat{\mathrm{g}}(x) = \hat{\beta}_0 + \hat{\beta}_1 x \qquad (3.17)$$

Vlerësimi i variancës për këtë vlerësues jepet nga barazimi 3.18.

$$\widehat{\mathrm{Var}}[\hat{\mathrm{g}}(x)] = \widehat{\mathrm{Var}}[\,\hat{\beta}_0 + \hat{\beta}_1 x] = \widehat{\mathrm{Var}}(\hat{\beta}_0) + x^2\widehat{\mathrm{Var}}(\hat{\beta}_1) + 2x\widehat{\mathrm{cov}}(\hat{\beta}_0, \hat{\beta}_1) \qquad (3.18)$$

Intervali i besimit i bazuar në statistiken Wald për modelin logit jepet si më poshtë:

$$\hat{\mathrm{g}}(x) \pm z_{1-\alpha/2}\widehat{\mathrm{SE}}[\hat{\mathrm{g}}(x)] \qquad (3.19)$$

$$\text{ku } \widehat{\mathrm{SE}}[\hat{\mathrm{g}}(x)] = \sqrt{\widehat{\mathrm{Var}}[\hat{\mathrm{g}}(x)]}$$

3.3 Regresioni Logjistik i Shumfishtë

Në këtë paragraf do të përgjithsohet modeli logjistik i thjështë në rastin e më shumë se një variabëli të pavarur. Duke supozuar se janë p varaiabla të pavarur $x_{1,}x_{2,}\dots,x_p$ dhe një variabel i varur $Y = \{0, 1\}$.

Shënohet x' vektori me komponente këto variabla, $x' = (x_{1,}x_{2,}\dots,x_{p,})$.

Shënohet $P = (Y = 1/x) = \pi(x)$. në rastin kur të gjithë variablat e pavarur janë në shkallë intervalore, modeli Logit për modelin logjistik të shumëfishtë jepet nga ekuacioni 3.20:

$$g(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p \qquad (3.20)$$

kur modeli i regresit logjistik është:

$$\pi(x) = \frac{e^{g(x)}}{1+e^{g(x)}} \qquad (3.21)$$

Nëse disa nga variablat e pavarur janë kategorik në shkallën nominale ose të zakonshme, ata duhet të konsiderohen si variabla të dizenjuar. Në përgjithësi, nëse një variabël me shkallë nominale ka *k* vlera të mundshme, atëherë ndërtohen $k-1$ variabla të dizenjuara. Për ta ilustruar ndertimin e variabli të dizenjuar supozojmë se variabli i pavaraur x_j ka k_j vlera. Për variablin x_j do të ndërtohen $k_j - 1$ variabla të dizenjuara të cilët shënohen D_{jl} për $l = 1,2,\dots,k_j - 1$, me koefiçentë përkates β_{jl}. Kështu modeli logit 3.20 me p variabla ku variabli j^{-te} është kategorik merr trajten:

$$g(x) = \beta_0 + \beta_1 x_1 + \cdots + \sum_{l=1}^{k_j-1} \beta_{jl} D_{jl} + \beta_p x_p \qquad (3.22)$$

3.3.1 Përshtatja e Modelit të Regresit Logjistik të Shumëfishtë

Supozohet një zgjedhje me *n* vrojtime të pavarura (x_i, y_i), *i= 1,2,…, n.* Njësoj si në rastin e regresit të regresit të thjështë logjistik, përshtatja e modelit kërkon gjetjen e vlerësueseve të parametrave në vektorin $\beta' = (\beta_0, \beta_1, \dots, \beta_p)$. Metoda e vlerësimit do të ajo e përgjasise maksimale. Funksioni i përgjasise është thuajse identik me ekuacionin e dhënë (3.3) me ndryshimin e vetëm se $\pi(x)$ është ai i përcaktuar në ekuacionin 3.21. Janë $p+1$ ekuacione përgjasie të cilët përftohen duke llogaritur derivatet e pjesshme të funksionit të përgjasise në lidhje me p+1 koeficientet respektive. Ekuacionet e përgjasise jepen në barazimet e meposhtëme:

$$\sum_{i=1}^{n} [y_i - \pi(x_i)] = 0$$

dhe

$$\sum_{i=1}^{n} x_{ij}[y_i - \pi(x_i)] = 0$$

Për *j= 1, 2…, p.*

Zgjidhja e ekuacioneve të përgjasisë bëhet nëpërmjet programeve të posaçëm qe janë të disponueshme thuajse në të gjitha paketat statistikore. Me $\hat{\beta}$ shënohet zgjidhja e ketyre ekuacioneve. Vlerat e përshtatura për modelin e regresit logjistik të shumëfishte janë $\hat{\pi}(x_i)$, vlera në ekuacionin 3.21 llogaritet duke përdorur $\hat{\beta}$, dhe x_i.

Metoda për vlerësimin e variancës dhe kovariances se vlerësueseve të koeficentëve ndjek parimin e përgjasisë maksimale. Vlerësuesit merren nga matricat e derivateve të pjesshme të rendit të dytë të funksionit log të përgjasisë. Këto derivate të pjesshme të rendit të dytë jepen nga barazimet 3.23 dhe 3.24:

$$\frac{\partial^2 L(\beta)}{\partial \beta_j^2} = -\sum_{j=1}^{n} x_{ij}^2 \pi_i (1 - \pi_i) \qquad (3.23)$$

dhe

$$\frac{\partial^2 L(\beta)}{\partial \beta_j \partial \beta_l} = -\sum_{i=1}^{n} x_{ij} x_{il} \pi_i (1 - \pi_i) \qquad (3.24)$$

Për *j, l = 0,1,2,…,p.* ku $\pi_i = \pi(x_i)$. Shënohet me $I(\beta)$ matrica me $(p+1)$x$(p+1)$ përmasa e cila ka për elemente termat negative të ekuacioneve 3.23 dhe 3.24. Kjo matricë, quhet matrica e informacionit të vrojtuar. Variancat dhe kovariancat e vlerësuesve të koefiçentëve gjenden nga matrica e anasjelle e saj, $Var(\beta) = I^{-1}(\beta)$. Shënojmë $Var(\beta_j)$ elementin e *j-te* të diagonales të kësaj matrice, e cila është varianca e $\hat{\beta}_j$, dhe shenojme $Cov(\beta_j, \beta_l)$ elementët jashtë diagonales kovariancën midis $\hat{\beta}_j$ dhe $\hat{\beta}_l$. Vlerësuesit e variancës dhe kovariancës të cilët shënohen me $\widehat{Var}(\hat{\beta})$, merren duke vlerësuar $Var(\beta)$ për $\hat{\beta}$. Shënojme me $\widehat{Var}(\hat{\beta}_j)$, dhe $\widehat{Cov}(\hat{\beta}_j, \hat{\beta}_l), j, l = 0,1,2 \dots, p$ elementet e matricës.

Vleresuesi i gabimit standard për vlerësuesit e koeficienteve jepet në barazimin3.25:

$$\widehat{SE}(\hat{\beta}_j) = \sqrt{\widehat{Var}(\hat{\beta}_j)} \qquad (3.25)$$

për $j= 0,1,2,\ldots,p$.

Një trajtë matricore e matricës se informacionit të vrojtuar është $\hat{I}(\hat{\beta}) = X'VX$, ku X është një matricë me *n* x (*p+1)* përmasa që përmban të dhëna për secilin vrojtim dhe V është një matrice me përmasa *nxn* me elementë të trajtes $\hat{\pi}_i(1-\hat{\pi}_i)$.

Pra:

$$X = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}$$

$$V = \begin{bmatrix} \hat{\pi}_1(1-\hat{\pi}_1) & 0 & \cdots & 0 \\ 0 & \hat{\pi}_2(1-\hat{\pi}_2) & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & \hat{\pi}_n(1-\hat{\pi}_n) \end{bmatrix}$$

3.3.2 Kontrolli i Rëndësisë së Modelit

Hapi i parë për vlerësimin e modelit të regresit logjistik të shumëfishtë, njësoj si në rastin e regresit të thjeshtë logjistik është vlerësimi i rëndësisë se variablave në model. Kriteri i raportit të përgjasisë për rëndësinë e përgjithshme të *p* koefiçentët e variablave të pavarur në model zbatohet në të njëjtën mënyrë si në rastin e modelit të regresit të thjështë logjistik. Kriteri bazohet në statistiken G të dhënë në ekuacionin 3.11 me ndryshimin qe vlerat e përshtatura, $\hat{\pi}$, në model bazohen në vektorin $\hat{\beta}$ me p+1 parametra.

Lidhur me hipotezën $H_0: \beta_1 = \beta_2 = \cdots = \beta_p = 0$, shpërndarja e statistikës *G* është Hi-katror me *p* shkallë lirie.

Për kontrollin e rëndësisë të secilit variabël të pavarur më vete përdoret statistika Wald. Për hipotezën $H_0: \beta_j = 0$ statistika Wald ka trajtën : $W_j = \frac{\hat{\beta}_j}{\widehat{SE}(\hat{\beta}_j)}$, secila statistike W_j, ka ligj normal standard. Qëllimi është që të merret modeli më i përshtatshëm ndërkohë që reduktohet numri i parametrave.

Hapi tjetër është të përshtatet modeli i reduktuar i cili përmban vetëm variablat e rëndësishëm dhe krahasimi i tij me modelin qe përmban të gjithë variablat. Duhet kujdes në zbatimin e statistikës Wald për rëndësin e secilit parametër për shkak të numrit të madh të shkallëve të lirise.

Për kontrollin e hipotezes H_0: $\beta_0 = \beta_1 = \cdots = \beta_p = 0$, kriteri Wald analog i modelit të shumëfishtë përftohet në trajtë matricore si me poshtë:

$$W = \hat{\beta}'[\widehat{Var}(\hat{\beta}_j)]^{-1}\hat{\beta} = \hat{\beta}'(X'VX)\hat{\beta}$$

e cila ka shpërndarje Hi-katror me *p+1* shkallë lirie.

3.4 Regresi Logjistik Multinomial

Në paragrafet e mësipërm u përshkrua modeli i regresit logjistik kur variabli i varur është vetëm me dy vlera (binar). Modeli mund të modifikohet thjeshtë për të modeluar rastin kur variabli i varur është nominal me më shumë se dy kategori. Në këtë rast supozohet një zgjerim i modelit të regresit logjistik i cili quhet modeli i regresit logjistik multinominal.

Në modelin e regresit logjistik multinomial duhet bërë kujdes për shkallën e matjes të variablit të varur i cili përgjithësisht është shkallë nominale, rast i cili do të shtjellohet në këtë paragraf por ka edhe raste kur shkalla e matjes se variablit të varur është e zakonshme.

Supozohet se variabli i varur *Y* ka *k* kategori të koduara *0,1,2,...,k-1*. Në modelin e regresit logjistik multinomial duhen ndërtuar *k-1* funksione logit të cilët do të ndërtohen duke caktuar paraprakisht një kategori reference (bazë) krahasimi. Zgjerimi më i qartë i modelit bëhet duke caktuar $Y = 0$ si kategori reference dhe formohen funksionet logit krahasues për $Y = 1$, $Y = 2$, ..., $Y = k - 1$.

Funksioni logit i $Y = r$ përkundrejt $Y = r - 1\ per\ r = 2,3, \dots, k - 1$ është diferenca ndërmjet këtyre dy funksioneve logit.

Supozohet se janë *p* variabla të varur komponente të vektorit x dhe një term konstant, $x_0 = 1$. Shënojmë funksionet logit si:

$$g_1(x) = ln\left[\frac{P(Y = 1 \,/\, x)}{P(Y = 0 \,/\, x)}\right] = \beta_{10} + \beta_{11}x_1 + \beta_{12}x_2 + \cdots + \beta_{1p}x_p = x'\beta_1$$

dhe

$$g_2(x) = ln\left[\frac{P(Y = 2 \,/\, x)}{P(Y = 0 \,/\, x)}\right] = \beta_{20} + \beta_{21}x_1 + \beta_{22}x_2 + \cdots + \beta_{2p}x_p = x'\beta_2$$

.

.

.

$$g_{\mathrm{k}-1}(x) = ln\left[\frac{P(Y=\mathrm{k}-1\ /\ x)}{P(Y=0\ /\ x)}\right] = \beta_{\mathrm{k}-1\,0} + \beta_{\mathrm{k}-1\,1}x_1 + \beta_{\mathrm{k}-1\,2}x_2 + \cdots + \beta_{2p}x_p = x'\beta_{\mathrm{k}-1}$$

Rrjedh se probabilitet e kushtëzuara të çdo kategorie për vektorin e varur x të janë:

$$P(Y=0/x) = \frac{1}{1+e^{g_1(x)}+e^{g_2(x)}+\cdots+e^{g_{\mathrm{k}-1}(x)}} \qquad (3.26)$$

$$P(Y=1/x) = \frac{e^{g_1(x)}}{1+e^{g_1(x)}+e^{g_2(x)}+\cdots+e^{g_{\mathrm{k}-1}(x)}} \qquad (3.27)$$

...

$$P(Y=\mathrm{k}-1/x) = \frac{e^{g_{\mathrm{k}-1}(x)}}{1+e^{g_1(x)}+e^{g_2(x)}+\cdots+e^{g_{\mathrm{k}-1}(x)}} \qquad (3.28)$$

Duke shënuar $\pi_j(x) = P(Y = j\ /\ x)$ për $j = 0,1,2,\dots,k-1$ çdo probabilitet është një funksion i vektorit të $2(p+1)$ parametrave $\beta' = (\beta_1', \beta_2', \dots, \beta_k')$.

Një shprehje e përgjithshme për probabilitetin e kushtëzuar në modelin me *k* kategori është:

$$P(Y=j/x) = \frac{e^{g_j(x)}}{\sum_{j=0}^{\mathrm{k}-1} e^{g_j(x)}}$$

$$\text{ku} \quad g_{\mathrm{j}}(x) = ln\left[\frac{P(Y=\mathrm{j}/x)}{P(Y=0/x)}\right] = \beta_{\mathrm{j}\,0} + \beta_{\mathrm{j}\,1}x_1 + \beta_{\mathrm{j}\,2}x_2 + \cdots + \beta_{\mathrm{j}p}x_p = x'\beta_{\mathrm{j}}$$

Kur vektorët $\beta_0 = 0$ dhe $g_0(x) = 0$ parametrat vlerësohen njësoj si në rastin e regresit logjistik, nëpërmjet parimit të përgjasise maksimale.

3.5 Modelet e regresit logjistik për identifikimin e autorit

3.5.1 Regresi i shumëfishtë logjistik

Në këtë paragraf paraqitet modeli i regresit logjistik të shumëfishtë për identifikimin e autorit të teksteve.

Duke shënuar X vektorin me elemente të gjitha karakteristikat gjuhësore e sintaksore të një teksti dhe Y vektorin që merr vetëm dy vlera 0 nëse teksti nuk është i autorit të caktuar dhe 1 nëse teksti është i autorit të caktuar. Për modelimin e këtij problem përdoret modeli i regresit logjistik me shumë variabla të pavarur dhe një variabël binar të pavarur.

Duke shënuar: $X = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$, funksioni logjistik merr trajten:

$$f(X) = \frac{1}{1+e^{-(\alpha+\sum_i \beta_i x_i)}} = \frac{e^{(\alpha+\sum_i \beta_i x_i)}}{1+e^{(\alpha+\sum_i \beta_i x_i)}}\,.$$

Për të llogaritur probabilitetin *P(Y=1/X)* që teksti me karakteristika X të jetë shkruar nga autori i caktuar me ndihmën e funksionit logjistik shkruaj:

$$P(Y=1|X)=\frac{e^X}{1+e^X}=\frac{e^{(\alpha+\sum_i \beta_i x_i)}}{1+e^{(\alpha+\sum_i \beta_i x_i)}}$$

Vlerësimi i parametrave të panjohur α e β_i bëhet me metodën e përgjasisë maksimale.

Në qoftë se shënoj p probabilitetin që teksti të jetë i autorit të caktuar atëherë sipas modelit të regresit logjistik:

$$\frac{p}{1-p}=\frac{\frac{e^X}{1+e^X}}{1-\frac{e^X}{1+e^X}}=e^X => \log\left(\frac{p}{1-p}\right)=\alpha+\sum_i \beta_i x_i \ , ku\ p=P(Y=1|X).$$

Duke përdorur transformimin logjistik shkruaj:

$$P(Y=1|X)=\frac{e^X}{1+e^X}=\frac{e^{(\alpha+\sum_i \beta_i x_i)}}{1+e^{(\alpha+\sum_i \beta_i x_i)}}=\frac{e^{\alpha}e^{\sum_i \beta_i x_i}}{1+e^{\alpha}e^{\sum_i \beta_i x_i}}=\frac{e^{\alpha}e^{\sum_i \beta_i x_i}}{e^{\alpha}(e^{-\alpha}+e^{\sum_i \beta_i x_i})}=\frac{e^{\sum_i \beta_i x_i}}{(e^{-\alpha}+e^{\sum_i \beta_i x_i})}.$$

3.5.2 Regresi logjistik multinomial

Le të jenë: $x=[x_1,\dots,x_j,\dots,x_d]^T$ vektori me elemente karakteristikat e një teksti të cilit do ti përcaktohet autoresia. $y=[y_1,\dots,y_K]^T$, për secilin tekst y_k=1 në qoftë se teksti është i autorit k dhe y_i=0 (Ose $y=[y_1,\dots,y_K]^T$ vektor me aq elemente sa tekste janë në studim, me k vlera (kategori) të ndryshme. Për tekstet e autorit k, përkatësisht koordinatat e *y* shënohen k.).

$B=[\beta_1,\dots,\beta_K]^T$ matrica e parametrave të modelit, ku $\beta_k=[\beta_{k1},\dots,\beta_{kd}]^T$ parametri për autorin *k*. $P(y_k=1\mid x,B)=\frac{\exp(\beta_k^T x)}{\sum_{k'}\exp(\beta_{k'}^T x)}$, modeli i probabilitetit me kusht.

Autori me probabilitetin me të madh do të jetë autori i vërtetë i tekstit.

$$\hat{y}(x)=arg\ \underbrace{\arg max}_{k}\ P(y_k=1\mid x)\ .$$

Në qoftë se shënohet D bashkësia e teksteve dhe autoreve të cilët shqyrtohen. $D=\{(x_1,y_1),\dots,(x_i,y_i),\dots,(x_n,y_n)\}$, bashkësia e parametrave B e llogaritet nëpërmjet përgjasise maksimale duke vlerësuar:

$$l(B|D)=-\sum_i[\sum_k y_{ik}\beta_k^T x_i-ln\sum_k \exp(\beta_k^T x_i)]$$

3.6 Modeli i regresit logjistik për tekste në gjuhën shqipe

3.6.1 Variablat për tekstet në gjuhën shqipe

Ndër tekstet shqip që u shqyrtuan në studim në kapitullin e dytë janë 7 romane të 3 autoreve të ndryshëm shqiptare me mesatarisht 59924 fjale dhe 264242 germa secili roman dhe 100 artikuj të 10 autorëve të ndryshëm me mesatarisht 1280 fjalë dhe 6216 germa. Për secilin tekst u vrojtuan: numri i fjalëve, numri i germave, numri i zanoreve, numri i bashkëtingëlloreve, numri i fjalive, numri i shenjave të pikësimit, frekuenca e lidhëzës *se* dhe frekuenca e lidhëzës *që*. Në këtë kapitull këto vrojtime konsiderohen si variabla të pavarur për të ndërtuar modele të regresit logjistik për tekstet shqip. Meqenëse tekstet ndryshojnë për nga madhësia dhe grupi i romaneve është i vogël, do të ndërtohen modele të regresit logjistik vetëm për 100 artikujt.

Më poshtë emërtojmë variablat duke shënuar:

NrFjale = numrin e fjalëve,

NrGerma = numrin e germave,

NrZanore = numrin e zanoreve,

NrBashketingellore = numrin e bashkëtingëlloreve,

NrFjali = numrin e fjalive,

NrShenjapikesimi = numrin e shenjave të pikësimit,

NrSe = frekuencen e lidhëzës *se* ,

NrQE = frekuencën e lidhëzës *që*,

G_F = NrGerma/ NrFjale,

B_Z= NrBashketingellore/ NrZanore.

AUTOR*J*=vektor shtylle me 100 elemente 0 dhe 1. Për secilin autor do të këtë 0 për tekstet të cilët nuk janë të autorit *J* dhe 1 për tekstet qe janë të autorit *J*. Janë 10 variabla qe konsiderohen të varur për modelet e regresit logjistik të shumëfishtë pra

$J \epsilon \{1,2,3,4,5,6,7,8,9,10\}$.

AUTORI= vektor shtylle me 100 elemente numrat 1, 2, ..., 10. Do të shenohet 1 nese teksti është i autorit 1, 2 nese teksti është i autorit 2 edhe keshtu me radhe. Ky do të konsiderohet si variabel i varur për modelet e regresit logjistik multinomial.

3.6.2 Modelet e regresit logjistik të shumëfishtë për tekset shqip

Në punimin Salillari et al. (2012) kemi ndërtuar një model të regresit logjistik për 6 romane, duke konsideruar vetëm numrin e fjalëve, numrin e germave dhe numrin e

zanoreve. Parametrat e vlerësuar në atë model nuk ishin më të mirët për shkak të numrit të vogël të variablave dhe teksteve. Meqenëse tekstet shqip i kemi të madhësive të ndryshme romane dhe artikuj gazetash, do ndërtojmë modele të regresit logjistik vetëm për 100 artikujt duke shtuar numrin e variablave shpjegues.

Në punimin Salillari.D, Prifti (2016) kemi ndërtuar modele të regresit të shumëfishtë logjistik për 100 tekstet duke konsideruar si variabla të pavarur numrin e fjalëve (NrFjale), numrin e germave(NrGerma), numrin e zanoreve (NrZanore), numrin e bashketingëlloreve (NrBashketingellore), numrin e fjalive (NrFjali) dhe numrin e shenjave të pikësimit (NrShenjapikesimi) për secilin tekst dhe si variabla të varur variablat AUTOR*J*.

Modeli fillestar qe morëm ishte:

```
model<-glm(Author1~N_words+N_letters+N_vowels+N_consonant+N_sentences+N_punctuations,
data = paper,family= binomial())
```

```
> summary(model1,corr=T)

Call:
glm(formula = Author1 ~ N_words + N_letters + N_vowels + N_consonant +
    N_sentences + N_punctuations, family = binomial(), data = paper)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4186  -0.2095  -0.1057  -0.0315   3.5414

Coefficients: (1 not defined because of singularities)
                Estimate Std. Error z value Pr(>|z|)
(Intercept)    -0.102289   1.761839  -0.058  0.95370
N_words        -0.011986   0.011585  -1.035  0.30085
N_letters       0.006257   0.003962   1.579  0.11428
N_vowels       -0.004237   0.007422  -0.571  0.56810
N_consonant           NA         NA      NA       NA
N_sentences     0.145752   0.045147   3.228  0.00124 **
N_punctuations -0.162807   0.054598  -2.982  0.00286 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 65.017  on 99  degrees of freedom
Residual deviance: 29.031  on 94  degrees of freedom
AIC: 41.031

Number of Fisher Scoring iterations: 7
```

```
Correlation of Coefficients:
              (Intercept) N_words N_letters N_vowels N_sentences
N_words       -0.28
N_letters      0.06       -0.57
N_vowels       0.07        0.10   -0.85
N_sentences   -0.17        0.15    0.09     -0.01
N_punctuations  0.00       -0.26   -0.14      0.10    -0.85
```

Nga rezultatet e këtij modeli u vërejt se vetëm dy prej variablave ishin statistikisht të rëndësishëm dhe se ka një korrelim të lartë midis numrit të zanoreve dhe numrit të bashketingëlloreve.

Pas shqyrtimit të modelve të ndryshme modeli me i mirë qe u përftua ishte modeli 1 si më poshtë:

```
> Model1 <- glm(Author1 ~ N_words+N_letters+N_sentences+N_punctuations, data = paper,family = binomial())
> summary(Model1,corr=T)

Call:
glm(formula = Author1 ~ N_words + N_letters + N_sentences + N_punctuations,
    family = binomial(), data = paper)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4454  -0.2173  -0.1020  -0.0303   3.5532

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)     0.062957   1.725130   0.036  0.97089
N_words        -0.011299   0.011095  -1.018  0.30847
N_letters       0.004416   0.002038   2.166  0.03028 *
N_sentences     0.149242   0.045383   3.288  0.00101 **
N_punctuations -0.165346   0.054129  -3.055  0.00225 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 65.017  on 99  degrees of freedom
Residual deviance: 29.402  on 95  degrees of freedom
AIC: 39.402
```

```
Number of Fisher Scoring iterations: 7

Correlation of Coefficients:
              (Intercept) N_words N_letters N_sentences
N_words        -0.30
N_letters       0.23       -0.91
N_sentences    -0.18        0.14    0.18
N_punctuations  0.01       -0.26   -0.14     -0.86
```

Variablat shpjegues me të rëndësishëm për nivel rendesie $\alpha=0.05$ rezultuan numri i fjalëve, numri i germave, numri i fjalive dhe numri i shenjave të pikesimit.

Sipas rezultateve të përftuara në R modeli shkruhet:

$$\text{logitP(y)} = \ln\left[\frac{P(Y=1)}{1-P(Y=1)}\right] = 0.062957 - 0.011299\text{N}_{\text{words}} + 0.004416\text{N}_{\text{letters}} - 0.165346\text{N}_{\text{punctuations}} + 0.149242\text{N}_{\text{sentences}}$$

Kontrolloj autoresinë e tekstit të pestë Tekst15:

$P(Y=1|X_5) = 9.660309e-01$, $\text{P(Y} = 0|\text{X}_5) = 1 - 9.660309e-01 = 0.339698\text{e} - 01$

Vërej se

$$P(Y=1|X_5) > \text{P(Y} = 0|\text{X}_5)$$

Pra tekstit Tekst15 i caktohet si autor Autor 1.

Duke llogaritur probabilitetet për caktimin e autorësisë të gjithë teksteve u përftua tabela e mëposhtëme:

Teksti		Parashikimi I autoresise		
		AUTOR1		Përqindja e klasifikimit të sakte
		0	1	
AUTOR1	0	89	1	98.9
	1	3	7	70.0
Total		92	8	96.0

TABELA 3. 2: REZULTATET E KLASIFIKIMIT TE TEKSTEVE NE MODELIN 1

Për këtë model autorësia e teksteve identifikohet saktë me probabilitet 0.96 sipas rezultateve në tabelën 3.2. Si rezultat i shqyrtimit të modeleve për të gjithë autorët klasifikimi i saktë i teksteve u bë me probabilitet 0.918.

Në kapitullin 2 u shqyrtuan disa elementë të rëndësishëm sintaksor në gjuhën shqipe si lidhëzat "*që*", "*se*", koeficientet e raportit të bashketingëlloreve me zanoret edhe koeficientin e raportit të germave me fjalet në gjuhën shkruar. Për këtë arsye do të shtohen variabla të rinj për të përmiresuar modelet e regresit logjistik me qellim identifikimin e autoresise se teksteve.

Më poshtë për ndërtimin e modeleve fillimisht konsiderohen si variabla të pavarur numri i fjalëve, numri i germave, numri i bashketingëlloreve, numri i zanoreve, numri i fjalive, numri i shenjave të pikesimit, frekuenca e lidhëzës "që" dhe frekuenca e lidhezes "se".

```
> DATAARTSH<-read.table(file="clipboard",header=TRUE)

> LM1ARTA1 <- glm(AUTOR1 ~ NrFjale + NrGerma+NrBashketingellore+NrZanore+NrShenjapikesimi
+NrFjali+NrSE+NrQe, data = DATAARTSH,family = binomial())

> summary(MNLART)


Call:

glm(formula = AUTOR1 ~ NrFjale + NrGerma + NrBashketingellore +

    NrZanore + NrShenjapikesimi + NrFjali + NrSE + NrQe, family = binomial(),

    data = DATATSHQIP)


Deviance Residuals:

     Min        1Q    Median        3Q       Max

-1.41947  -0.00001   0.00000   0.00000   1.43603


Coefficients: (1 not defined because of singularities)
```

```
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)       -4.15573    7.29105  -0.570    0.569
NrFjale           -0.15055    0.20849  -0.722    0.470
NrGerma            0.01357    0.01564   0.867    0.386
NrBashketingellore 0.03879    0.05794   0.669    0.503
NrZanore                NA         NA      NA       NA
NrShenjapikesimi  -0.92909    1.07931  -0.861    0.389
NrFjali            1.09335    1.29182   0.846    0.397
NrSE              -1.33826    1.61405  -0.829    0.407
NrQe               1.50387    1.82103   0.826    0.409

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 65.017  on 99  degrees of freedom
Residual deviance: 12.060  on 92  degrees of freedom
AIC: 28.06

Number of Fisher Scoring iterations: 13
```

Nga rezultatet e modelit vërehet se asnjë nga koeficientet e modelit nuk del statistikisht i rëndësishëm për nivel rëndësie α të paktën 0.1. Gjithashtu vërehet se koeficienti për NrZanore nuk është vlerësuar. Kjo ndodh për shkak se kemi një korrelacion të lartë midis variablave siç edhe u vlerësua në kapitullin 2.

Më poshtë shqyrtohen modele ku reduktohen variabla. Në modelin e ri përjashtohen variablat NrZanore dhe NrBashketingellore dhe përftohet modeli i dytë për autorin 1 si më poshtë:

```
> LM2ARTA1 <- glm(AUTOR1 ~ NrFjale + NrGerma+NrShenjapikesimi+NrFjali+NrSE+NrQe, data = D
ATAARTSH,family = binomial())
> summary(LMARTA1)

Call:
glm(formula = AUTOR1 ~ NrFjale + NrGerma + NrShenjapikesimi +
    NrFjali + NrSE + NrQe, family = binomial(), data = DATATSHQIP)
```

```
Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.39661  -0.00796  -0.00102  -0.00003   1.41358

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)      -0.490662   2.208140  -0.222   0.8242
NrFjale          -0.050346   0.034487  -1.460   0.1443
NrGerma           0.013183   0.007464   1.766   0.0774 .
NrShenjapikesimi -0.425846   0.224086  -1.900   0.0574 .
NrFjali           0.494342   0.253707   1.948   0.0514 .
NrSE             -0.576399   0.428231  -1.346   0.1783
NrQe              0.641594   0.349489   1.836   0.0664 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 65.017  on 99  degrees of freedom
Residual deviance: 13.613  on 93  degrees of freedom
AIC: 27.613

Number of Fisher Scoring iterations: 11
```

Nga rezultatet e këtij modeli vërehet se të gjithë parametrat përveç parametrit për NrSE dhe për NrFjale, janë statistikisht të rëndësishëm me nivel rendesie $\alpha=0.1$.

Atëherë në modelin tjetër përjashtojme variablin NrSE meqenëse ka edhe vleren e p me të vogël.

```
> LM3ARTA1<- glm(AUTOR1 ~ NrFjale + NrGerma+NrShenjapikesimi+NrFjali+NrQe, data = DATAART
SH,family = binomial())
> summary(LMARTA12)
```

```
Call:
glm(formula = AUTOR1 ~ NrFjale + NrGerma + NrShenjapikesimi +
    NrFjali + NrQe, family = binomial(), data = DATATSHQIP)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.94919  -0.06007  -0.01581  -0.00150   1.71156

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)      -1.729221   1.891433  -0.914   0.3606
NrFjale          -0.042219   0.023987  -1.760   0.0784 .
NrGerma           0.010143   0.004703   2.157   0.0310 *
NrShenjapikesimi -0.251014   0.103355  -2.429   0.0152 *
NrFjali           0.287150   0.111651   2.572   0.0101 *
NrQe              0.300488   0.130997   2.294   0.0218 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 65.017  on 99  degrees of freedom
Residual deviance: 17.171  on 94  degrees of freedom
AIC: 29.171

Number of Fisher Scoring iterations: 9
```

Nga rezultatet e modelit vërehet se të gjithë parametrat tani janë statistikisht të rëndësishëm për nivel rendesie $\alpha=0.05$.

Për Autorin 1 ky model konsiderohet si modeli më i mirë.

Pasi vleresohen probabilitetet përftohet tabela për klasifikimin e teksteve si më poshtë.

		Parashikimi I autoresise		
		AUTOR1		Përqindja e klasifikimit të sakte
Teksti		0	1	
AUTOR1	0	87	3	96.7
	1	2	8	80.0
Total		89	11	95.0

TABELA 3. 3: REZULTATET E KLASIFIKIMIT NE MODELIN 3 TE AUTORIT 1

Pasi kemi shqyrtohen modelet për të gjithë autorët konstatohet se parametrat statistikisht të rëndesishëm nuk janë gjithnjë të njëjtë për secilin autor. Të gjithë modelet për secilin autor janë paraqitur në Shtojcën 2 (CD). Modeli 3 edhe pse jo për të gjithë autoret ishte më i miri ai klasifikoi autorësinë e teksteve me një probabilitet 0.93. Rezultatet e klasifikimit të teksteve në secilin model jepen më poshtë në tabelën 3.4:

AUTORI	Klasifikimi i sakte në %
1	95
2	92
3	100
4	100
5	90
6	93
7	89
8	89
9	90
10	92
TOTAL	93

TABELA 3. 4: REZULTATET E KLASIFIKIMIT TE MODELIT 3 PER SECILIN AUTOR

Pra si rezultat i shtimit të dy variablave të tjerë kemi një rritje të probabilitetit të klasifikimit të saktë të teksteve nga 0.918 në 0.93.

Meqenëse variablat G_F dhe B_Z formohen si raporte përkatësisht G_F = NrGerma/ NrFjale dhe B_Z= NrBashketingellore/ NrZanore, me poshtë do të shqyrtohet çfarë

ndodh në model kur perjashtohen variablat NrGerma, NrFjale, NrBashketingellore, NrZanore dhe i zëvendësojme ato me variablat G_F dhe B_Z.

```
> LM4ARTA1<- glm(AUTOR1 ~ G_F+NrShenjapikesimi+NrFjali+NrQe+B_Z, data = DATAARTSH,family
= binomial())
> summary(LM4ARTA1)

Call:
glm(formula = AUTOR1 ~ G_F + NrShenjapikesimi + NrFjali + NrQe +
    B_Z, family = binomial(), data = DATAARTSH)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.90752  -0.02172  -0.00321  -0.00016   1.43480

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)      -139.02866   67.93630  -2.046   0.0407 *
G_F                26.05596   12.76969   2.040   0.0413 *
NrShenjapikesimi   -0.19661    0.08069  -2.437   0.0148 *
NrFjali             0.27498    0.11435   2.405   0.0162 *
NrQe                0.39263    0.16518   2.377   0.0175 *
B_Z                 4.82772   11.22262   0.430   0.6671
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 65.017  on 99  degrees of freedom
Residual deviance: 14.340  on 94  degrees of freedom
AIC: 26.34

Number of Fisher Scoring iterations: 10
```

Nga rezultatet e modelit kemi përcaktuar variablat G_F, B_Z, NrFjali, NrShenjapikesimi si edhe NrQE si variabla statistikisht të rëndësishëm për nivel rendesie α=0.05. Ky model

klasifikon saktë tekstet me probabilitet më të madh. Në tabelën 3.5 për autorin 1 probabiliteti i klasifikimit rritet me 2% krahasuar me modelin e mëparshëm.

Teksti		Parashikimi I autoresise		
		AUTOR1		Përqindja e klasifikimit të sakte
		0	1	
AUTOR1	0	88	2	97.8
	1	1	9	90.0
Total		89	11	97.0

TABELA 3. 5: REZULTATET E KLASIFIKIMIT TE TEKSTEVE NE MODELIN 4 TE AUTORIT 1

Pasi kemi shqyrtohen modelet për të gjithë autorët konstatohet njesoj si ne modelin 3 se parametrat statistikisht të rëndesishëm nuk janë gjithnjë të njëjtë për secilin autor. Të gjithë modelet për secilin autor janë paraqitur në Shtojcen 2 (CD). Modeli 4 edhe pse jo për të gjithë autoret ishte më i miri ai klasifikoi autorësinë e teksteve me një probabilitet 0.935. Rezultatet e klasifikimit të teksteve në secilin model janë përmbledhur më poshtë në tabelën 3.6:

AUTORI	Klasifikimi i sakte në %
1	97
2	93
3	100
4	100
5	91
6	93
7	89
8	90
9	90
10	92
TOTAL	93.5

TABELA 3. 6:REZULTATET E KLASIFIKIMIT TE MODELIT 4 PER SECILIN AUTOR

Për të modeluar tekstet me regresin e shumëfishtë logjistik duhet të ndërtojmë aq modele sa është numri i autorëve gje e cila kërkon më shumë kohë për identifikimin e autorësise së teksteve.

Problemi i klasifikimit të teksteve në një autor të caktuar mund të modelohet edhe me ndihmën e regresit logjistik multinomial. Shqyrtimi i këtyre modeleve do të bëhet në paragrafin që vijon.

3.7 Modeli i regresit logjistik multinomial për tekstet në gjuhën shqipe

Në punimin Salillari, Prifti (2016) u paraqit modeli i regresit multinomial për tektet shqip duke konsideruar si variabla të pavarur numrin e fjaleve, numrin e germave, numrin e zanoreve, numrin e bashketingelloreve, numrin e fjalive dhe numrin e shenjave të pikesimit. Rezultatet janë përftuar nëpërmjet SPSS. Modeli me i mirë në këtë punim rezultoi për pese variablat e pavarur duke përjashtuar numrin e germave. Vlerësimi i parametrave jepet në tabelën 3.7.

Likelihood Ratio Tests

Effect	Model Fitting Criteria	Likelihood Ratio Tests		
	-2 Log Likelihood of Reduced Model	Chi-Square	Df	Sig.
Intercept	259.230	52.751	9	.000
N_words	276.699	70.220	9	.000
N_consonant	262.179	55.699	9	.000
N_vowels	230.636	24.156	9	.004
N_punctuations	234.433	27.954	9	.001
N_sentences	249.452	42.973	9	.000

The chi-square statistic is the difference in -2 log-likelihoods between the final model and a reduced model. The reduced model is formed by omitting an effect from the final model. The null hypothesis is that all parameters of that effect are 0.

TABELA 3. 7: VLERESIMI I PARAMETRAVE NE MODELIN E PUNIMIT SALILLARI, PRIFTI (2016)

Për këtë model u përftua tabela e klasifikimit si në tabelën 3.8 nga e cila rezulton se klasifikimi i teksteve sipas këtij modeli ku të gjithë parametrat e tij janë statistikisht të rëndësishëm për α=0.05 behet me saktësi 0.59.

Classification

Observed	Predicted										
	1.00	**2.00**	**3.00**	**4.00**	**5.00**	**6.00**	**7.00**	**8.00**	**9.00**	**10.00**	**Percent Correct**
1.00	9	1	0	0	0	0	0	0	0	0	90.0%
2.00	1	7	0	0	0	0	0	2	0	0	70.0%
3.00	0	0	10	0	0	0	0	0	0	0	100.0%
4.00	0	0	0	6	2	1	0	0	0	1	60.0%
5.00	0	0	0	3	3	2	0	0	2	0	30.0%
6.00	0	1	0	2	1	3	0	0	2	1	30.0%
7.00	1	0	0	0	0	0	6	3	0	0	60.0%
8.00	1	1	0	0	0	0	3	4	0	1	40.0%
9.00	0	0	0	1	1	0	0	0	6	2	60.0%
10.00	0	0	0	1	0	2	1	0	1	5	50.0%
Overall Përcentage	12.0%	10.0%	10.0%	13.0%	7.0%	8.0%	10.0%	9.0%	11.0%	10.0%	59.0%

TABELA 3. 8: REZULTATET E KLASIFIKIMIT PER MODELIN NE PUNIMIN SALILLARI, PRIFTI (2016)

Në të dhënat tona vërehet prania e vlerave outlier të cilat u korrespondojnë teksteve të njëjta për gjithë variablat. Ndërtimi i modelit pa outliers rriti klasifikimin e teksteve me saktësi deri në 64%. Në tabelën 3.8 vërehet se përqindja më e ulët e klasifikimit të teksteve është për autorin e pestë dhe të gjashtë. Duke bërë krahasimet e shumëfishta për mesataret për të gjitha variablat ndërmjet autorëve, vërehet se të gjitha autorët janë të ndarë në dy grupe. Grupi i parë i autorëve 2, 3, 6, dhe 9 rezulton se ata nuk kanë ndryshime të rëndësishme midis mesatareve të variablave dhe grupin e dytë të autorëve 1,4,5,7,8 dhe 10 i në të cilin rezulton gjithashtu se ata nuk kanë asnjë ndryshim të rëndësishëm midis mesatareve të variablave. Meqenëse autori pestë dhe i gjashtë i përkasin grupeve të ndryshme, janë thuasje të njëjte me autorë të tjerë të grupeve përkatese dhe kanë përqindjen më të ulët në klasifikimin e teksteve, i përjashtuam nga studimi dhe shqyrtuam një model të ri i cili na rrit përqindjen e klasifikimit të sakte të teksteve në 73.8%. Në të gjitha modelet e regresit multinomial logjistik të shqyrtuar në këtë punim klasifikimi i teksteve të autorit të tretë është parashikuar me saktësi 100%. kjo

ndodh për shkak se ai ndryshe nga autorët e tjerë ai shkruan me fjali më të gjata dhe përdor më pak shenja pikësimi. Si konkluzion në këtë punim u përcaktuan parametra të ndryshem nga modeli i regresit të shumëfishtë logjistik te paraqitura ne Salillari.D, Prifti (2016).

Më poshtë po paraqesim të njëjtin model te diskutuar ne Salillari, Prifti (2016) por të realizuar në programin R:

```
> MLARTSH<-multinom(AUTORI~NrFjale+NrZanore+NrFjali+NrShenjapikesimi+NrBashketingellore,d
ata=DATAARTSH)
# weights:  70 (54 variable)
initial  value 230.258509
iter  10 value 210.956117
iter  20 value 178.099064
iter  30 value 167.970399
iter  40 value 164.663459
iter  50 value 112.494335
iter  60 value 105.495303
iter  70 value 104.024082
iter  80 value 103.798971
iter  90 value 103.374492
iter 100 value 103.255792
final  value 103.255792
stopped after 100 iterations

> summary(MLARTSH)

Call:

multinom(formula = AUTORI ~ NrFjale + NrZanore + NrFjali + NrShenjapikesimi +
    NrBashketingellore, data = DATAARTSH)

Coefficients:
   (Intercept)       NrFjale     NrZanore     NrFjali NrShenjapikesimi NrBashketingellore
2   -6.1417435 -0.0005959957 -0.003530348 -0.05546307        0.0858185        0.001526812
3  113.8110841  0.3228685960 -0.165489575 -2.68561681        0.3796405       -0.029152296
4    2.2733836  0.0707621456  0.025150678 -0.28558670        0.1907647       -0.045190876
5    5.1875309  0.0930897304 -0.019536146 -0.33126964        0.1517991       -0.021329853
6   -1.7356802  0.0776555288  0.007891552 -0.39362031        0.1850349       -0.032889658
7    5.4390698 -0.0083754349  0.004918121 -0.28740664        0.2168672       -0.006123472
8    4.2066381 -0.0162602645  0.006738492 -0.17685540        0.2260363       -0.006371113
9    9.3302508  0.0746985958 -0.009818288 -0.45259195        0.2321502       -0.024231181
10  -0.3766625  0.0508135124 -0.002715836 -0.41515617        0.2502058       -0.019075396

Std. Errors:
    (Intercept)    NrFjale    NrZanore     NrFjali NrShenjapikesimi NrBashketingellore
2  0.0022339937 0.01688709 0.003975929 0.051988487       0.06238440        0.004209214
3  0.0001259663 0.01596261 0.167603180 0.002154969       0.01075946        0.118813519
4  0.0005604314 0.02476597 0.014543939 0.089642406       0.07686337        0.011631391
5  0.0004229209 0.02562978 0.007000137 0.098373754       0.07805081        0.006036030
6  0.0017258283 0.02465512 0.011128378 0.103051279       0.07612111        0.010322578
```

```
7  0.0051111269 0.02267372 0.006953807 0.082470899       0.07035654        0.006627537
8  0.0043297780 0.01933100 0.007051889 0.066492119       0.07086731        0.006472403
9  0.0004541527 0.02572244 0.009578559 0.098451409       0.07756285        0.008174071
10 0.0004511709 0.02347172 0.008619999 0.095158324       0.07524413        0.007473490

Residual Deviance: 206.5116
AIC: 314.5116

> coeftest(MLARTSH)

z test of coefficients:


                         Estimate  Std. Error     z value  Pr(>|z|)
2:(Intercept)         -6.1417e+00  2.2340e-03  -2749.2215 < 2.2e-16 ***
2:NrFjale             -5.9600e-04  1.6887e-02     -0.0353 0.9718461
2:NrZanore            -3.5303e-03  3.9759e-03     -0.8879 0.3745782
2:NrFjali             -5.5463e-02  5.1988e-02     -1.0668 0.2860469
2:NrShenjapikesimi     8.5819e-02  6.2384e-02      1.3756 0.1689330
2:NrBashketingellore   1.5268e-03  4.2092e-03      0.3627 0.7168059
3:(Intercept)          1.1381e+02  1.2597e-04 903504.0482 < 2.2e-16 ***
3:NrFjale              3.2287e-01  1.5963e-02     20.2266 < 2.2e-16 ***
3:NrZanore            -1.6549e-01  1.6760e-01     -0.9874 0.3234519
3:NrFjali             -2.6856e+00  2.1550e-03  -1246.2437 < 2.2e-16 ***
3:NrShenjapikesimi     3.7964e-01  1.0759e-02     35.2843 < 2.2e-16 ***
3:NrBashketingellore  -2.9152e-02  1.1881e-01     -0.2454 0.8061763
4:(Intercept)          2.2734e+00  5.6043e-04   4056.4883 < 2.2e-16 ***
4:NrFjale              7.0762e-02  2.4766e-02      2.8572 0.0042735 **
4:NrZanore             2.5151e-02  1.4544e-02      1.7293 0.0837573 .
4:NrFjali             -2.8559e-01  8.9642e-02     -3.1858 0.0014433 **
4:NrShenjapikesimi     1.9076e-01  7.6863e-02      2.4819 0.0130696 *
4:NrBashketingellore  -4.5191e-02  1.1631e-02     -3.8853 0.0001022 ***
5:(Intercept)          5.1875e+00  4.2292e-04  12265.9602 < 2.2e-16 ***
5:NrFjale              9.3090e-02  2.5630e-02      3.6321 0.0002811 ***
5:NrZanore            -1.9536e-02  7.0001e-03     -2.7908 0.0052574 **
5:NrFjali             -3.3127e-01  9.8374e-02     -3.3675 0.0007586 ***
5:NrShenjapikesimi     1.5180e-01  7.8051e-02      1.9449 0.0517900 .
5:NrBashketingellore  -2.1330e-02  6.0360e-03     -3.5338 0.0004097 ***
6:(Intercept)         -1.7357e+00  1.7258e-03  -1005.7085 < 2.2e-16 ***
6:NrFjale              7.7656e-02  2.4655e-02      3.1497 0.0016345 **
6:NrZanore             7.8916e-03  1.1128e-02      0.7091 0.4782391
6:NrFjali             -3.9362e-01  1.0305e-01     -3.8197 0.0001336 ***
6:NrShenjapikesimi     1.8503e-01  7.6121e-02      2.4308 0.0150657 *
6:NrBashketingellore  -3.2890e-02  1.0323e-02     -3.1862 0.0014416 **
7:(Intercept)          5.4391e+00  5.1111e-03   1064.1625 < 2.2e-16 ***
7:NrFjale             -8.3754e-03  2.2674e-02     -0.3694 0.7118375
7:NrZanore             4.9181e-03  6.9538e-03      0.7073 0.4794074
7:NrFjali             -2.8741e-01  8.2471e-02     -3.4849 0.0004922 ***
7:NrShenjapikesimi     2.1687e-01  7.0357e-02      3.0824 0.0020534 **
7:NrBashketingellore  -6.1235e-03  6.6275e-03     -0.9239 0.3555156
8:(Intercept)          4.2066e+00  4.3298e-03    971.5598 < 2.2e-16 ***
8:NrFjale             -1.6260e-02  1.9331e-02     -0.8411 0.4002641
8:NrZanore             6.7385e-03  7.0519e-03      0.9556 0.3392953
8:NrFjali             -1.7686e-01  6.6492e-02     -2.6598 0.0078188 **
8:NrShenjapikesimi     2.2604e-01  7.0867e-02      3.1896 0.0014248 **
```

```
8:NrBashketingellore  -6.3711e-03  6.4724e-03     -0.9844 0.3249432
9:(Intercept)          9.3303e+00  4.5415e-04  20544.3052 < 2.2e-16 ***
9:NrFjale              7.4699e-02  2.5722e-02      2.9040 0.0036840 **
9:NrZanore            -9.8183e-03  9.5786e-03     -1.0250 0.3053502
9:NrFjali             -4.5259e-01  9.8451e-02     -4.5971 4.284e-06 ***
9:NrShenjapikesimi     2.3215e-01  7.7563e-02      2.9931 0.0027620 **
9:NrBashketingellore  -2.4231e-02  8.1741e-03     -2.9644 0.0030328 **
10:(Intercept)        -3.7666e-01  4.5117e-04   -834.8555 < 2.2e-16 ***
10:NrFjale             5.0814e-02  2.3472e-02      2.1649 0.0303967 *
10:NrZanore           -2.7158e-03  8.6200e-03     -0.3151 0.7527144
10:NrFjali            -4.1516e-01  9.5158e-02     -4.3628 1.284e-05 ***
10:NrShenjapikesimi    2.5021e-01  7.5244e-02      3.3253 0.0008834 ***
10:NrBashketingellore -1.9075e-02  7.4735e-03     -2.5524 0.0106981 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> predict(MLARTSH,DATAARTSH)
  [1] 1  1  1  1  1  2  1  1  1  1  2  2  2  2  2  8  1  2  8  2  3  3  3  3  3  3  3  3
 3  3  6  4  4  5  4
 [36] 4  4  10 5  4  4  9  4  5  6  5  5  6  9  4  6  4  6  4  10 6  9  2  5  9  8  1  7
 7  7  7  7  7  8  8
 [71] 8  8  2  8  10 7  7  7  8  1  9  4  9  9  9  5  9  10 10 9  10 4  10 7  10 10 9  10
 6  6
Levels: 1 2 3 4 5 6 7 8 9 10

> PredMult<-matrix(predict(MLARTSH11,DATAARTSH),nrow=10,byrow=TRUE)

> PredMult
      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
 [1,] "1"  "1"  "1"  "1"  "1"  "2"  "1"  "1"  "1"  "1"
 [2,] "2"  "2"  "2"  "2"  "2"  "8"  "1"  "2"  "8"  "2"
 [3,] "3"  "3"  "3"  "3"  "3"  "3"  "3"  "3"  "3"  "3"
 [4,] "6"  "4"  "4"  "5"  "4"  "4"  "4"  "10" "5"  "4"
 [5,] "4"  "9"  "4"  "5"  "6"  "5"  "5"  "6"  "9"  "4"
 [6,] "6"  "4"  "6"  "4"  "10" "6"  "9"  "2"  "5"  "9"
 [7,] "8"  "1"  "7"  "7"  "7"  "7"  "7"  "7"  "8"  "8"
 [8,] "8"  "8"  "2"  "8"  "10" "7"  "7"  "7"  "8"  "1"
 [9,] "9"  "4"  "9"  "9"  "9"  "5"  "9"  "10" "10" "9"
[10,] "10" "4"  "10" "7"  "10" "10" "9"  "10" "6"  "6"
```

Rezultatet e matricës PredMult i përmbledhim në tabelën 3.9.

AUTORI	Klasifikimi i sakte në %
1	90
2	70
3	100
4	60
5	30
6	30
7	60
8	40
9	60
10	50
Mesatarja	59

TABELA 3. 9: REZULTATET E KLASIFIKIMIT PER MODELIN E PARAQITUR NE PUNIMIN SALILLARI, PRIFTI(2016)

Me poshtë do të shqyrtohen modele të regresit multinomial logjistik duke shtuar variablat e pavarur NrQe, NrSE, G_F dhe B_Z .

```
> MLARTSH1<-multinom(AUTORI~NrFjale+NrGerma+NrZanore+NrBashketingellore+NrFjali+NrShenjap
ikesimi+NrQe+NrSE,data=DATAARTSH)
# weights:  100 (81 variable)
initial  value 230.258509
iter  10 value 211.964570
iter  20 value 154.561110
iter  30 value 140.073343
iter  40 value 132.717853
iter  50 value 116.278232
iter  60 value 98.245552
iter  70 value 65.224894
iter  80 value 51.269324
iter  90 value 47.949625
iter 100 value 45.696215
final  value 45.696215
stopped after 100 iterations

> summary(MLARTSH)
Call:
multinom(formula = AUTORI ~ NrFjale + NrGerma + NrZanore + NrBashketingellore +
    NrFjali + NrShenjapikesimi + NrQe + NrSE, data = DATAARTSH)

Coefficients:
   (Intercept)    NrFjale     NrGerma      NrZanore NrBashketingellore   NrFjali NrShenja
pikesimi      NrQe
```

```
2  -107.320847 0.03454881 -0.02306839 -0.0898857763        0.066817393 -2.359108
2.310180 -4.961512
3    70.070245 0.65938380 -0.10139050 -0.1111888842        0.009798371 -6.483285
2.076731 -3.034922
4   -33.830451 0.74420319 -0.08509576  0.0519603333       -0.137056090 -2.841723
2.094290 -5.991911
5     8.394002 0.53636250 -0.07730043 -0.0186860331       -0.058614382 -3.236288
2.123894 -3.593842
6   -51.265680 0.38446077 -0.03990164  0.1453352113       -0.185236808 -3.384206
2.026337 -3.383390
7    11.968003 0.38863804 -0.05799816 -0.0035409114       -0.054457233 -3.037939
2.155002 -3.691474
8    13.070534 0.36447036 -0.05597276 -0.0033781706       -0.052594559 -2.918725
2.173737 -3.756154
9    11.847652 0.50833953 -0.07380074 -0.0090475202       -0.064753189 -3.354404
2.218436 -3.470668
10   -2.838263 0.48515653 -0.06824254 -0.0009959675       -0.067246565 -3.318045
2.263889 -3.703334
          NrSE
2   8.34539669
3   4.95239022
4  -0.09109497
5   4.21395848
6   9.18424811
7   4.18147926
8   4.26760853
9   4.04186095
10  3.84892764

Std. Errors:
    (Intercept)    NrFjale     NrGerma   NrZanore NrBashketingellore    NrFjali NrShenjap
ikesimi        NrQe
2  1.513778e-03 0.06864966 0.009243338 0.02738155        0.023469751 0.21359408       0.1
2666215 0.071321034
3  8.519597e-04 0.10315449 0.015230524 0.05580870        0.048986560 0.03550310       0.0
9088060 0.139598750
4  5.892185e-05 0.08117061 0.011451441 0.01170947        0.001246694 0.01392261       0.0
2825309 0.003077289
5  2.777967e-03 0.06708952 0.009024402 0.01480275        0.016449561 0.07741799       0.0
7070405 0.088264222
6  2.847158e-03 0.04960220 0.008706015 0.07008431        0.064088492 0.18080864       0.0
6753637 0.178384579
7  2.530669e-03 0.07015524 0.009025290 0.01493301        0.016842013 0.08525863       0.0
7005774 0.103895588
8  3.086217e-03 0.07189965 0.009187970 0.01494350        0.016948673 0.08713132       0.0
6948574 0.105773905
9  3.110399e-03 0.06543091 0.008784744 0.01576373        0.017405979 0.08740995       0.0
6943443 0.084821517
10 5.782380e-03 0.06511719 0.008668562 0.01662177        0.018066945 0.07360091       0.0
6563745 0.111977216
          NrSE
2  0.042980891
3  0.061412581
4  0.002010941
5  0.138558810
```

```
6  0.077758591
7  0.136342159
8  0.148267251
9  0.122589108
10 0.166159756

Residual Deviance: 91.39243
AIC: 235.3924

> library(AER)
> coeftest(MLARTSH1)

z test of coefficients:

                        Estimate  Std. Error     z value  Pr(>|z|)
2:(Intercept)         -1.0732e+02  1.5138e-03 -7.0896e+04 < 2.2e-16 ***
2:NrFjale              3.4549e-02  6.8650e-02  5.0330e-01 0.6147796
2:NrGerma             -2.3068e-02  9.2433e-03 -2.4957e+00 0.0125717 *
2:NrZanore            -8.9886e-02  2.7382e-02 -3.2827e+00 0.0010281 **
2:NrBashketingellore   6.6817e-02  2.3470e-02  2.8470e+00 0.0044139 **
2:NrFjali             -2.3591e+00  2.1359e-01 -1.1045e+01 < 2.2e-16 ***
2:NrShenjapikesimi     2.3102e+00  1.2666e-01  1.8239e+01 < 2.2e-16 ***
2:NrQe                -4.9615e+00  7.1321e-02 -6.9566e+01 < 2.2e-16 ***
2:NrSE                 8.3454e+00  4.2981e-02  1.9417e+02 < 2.2e-16 ***
3:(Intercept)          7.0070e+01  8.5196e-04  8.2246e+04 < 2.2e-16 ***
3:NrFjale              6.5938e-01  1.0315e-01  6.3922e+00 1.635e-10 ***
3:NrGerma             -1.0139e-01  1.5231e-02 -6.6571e+00 2.794e-11 ***
3:NrZanore            -1.1119e-01  5.5809e-02 -1.9923e+00 0.0463358 *
3:NrBashketingellore   9.7984e-03  4.8987e-02  2.0000e-01 0.8414637
3:NrFjali             -6.4833e+00  3.5503e-02 -1.8261e+02 < 2.2e-16 ***
3:NrShenjapikesimi     2.0767e+00  9.0881e-02  2.2851e+01 < 2.2e-16 ***
3:NrQe                -3.0349e+00  1.3960e-01 -2.1740e+01 < 2.2e-16 ***
3:NrSE                 4.9524e+00  6.1413e-02  8.0641e+01 < 2.2e-16 ***
4:(Intercept)         -3.3830e+01  5.8922e-05 -5.7416e+05 < 2.2e-16 ***
4:NrFjale              7.4420e-01  8.1171e-02  9.1684e+00 < 2.2e-16 ***
4:NrGerma             -8.5096e-02  1.1451e-02 -7.4310e+00 1.078e-13 ***
4:NrZanore             5.1960e-02  1.1709e-02  4.4375e+00 9.103e-06 ***
4:NrBashketingellore  -1.3706e-01  1.2467e-03 -1.0994e+02 < 2.2e-16 ***
4:NrFjali             -2.8417e+00  1.3923e-02 -2.0411e+02 < 2.2e-16 ***
4:NrShenjapikesimi     2.0943e+00  2.8253e-02  7.4126e+01 < 2.2e-16 ***
4:NrQe                -5.9919e+00  3.0773e-03 -1.9471e+03 < 2.2e-16 ***
4:NrSE                -9.1095e-02  2.0109e-03 -4.5300e+01 < 2.2e-16 ***
5:(Intercept)          8.3940e+00  2.7780e-03  3.0216e+03 < 2.2e-16 ***
5:NrFjale              5.3636e-01  6.7090e-02  7.9947e+00 1.299e-15 ***
5:NrGerma             -7.7300e-02  9.0244e-03 -8.5657e+00 < 2.2e-16 ***
5:NrZanore            -1.8686e-02  1.4803e-02 -1.2623e+00 0.2068283
5:NrBashketingellore  -5.8614e-02  1.6450e-02 -3.5633e+00 0.0003663 ***
5:NrFjali             -3.2363e+00  7.7418e-02 -4.1803e+01 < 2.2e-16 ***
5:NrShenjapikesimi     2.1239e+00  7.0704e-02  3.0039e+01 < 2.2e-16 ***
5:NrQe                -3.5938e+00  8.8264e-02 -4.0717e+01 < 2.2e-16 ***
5:NrSE                 4.2140e+00  1.3856e-01  3.0413e+01 < 2.2e-16 ***
6:(Intercept)         -5.1266e+01  2.8472e-03 -1.8006e+04 < 2.2e-16 ***
6:NrFjale              3.8446e-01  4.9602e-02  7.7509e+00 9.126e-15 ***
6:NrGerma             -3.9902e-02  8.7060e-03 -4.5832e+00 4.579e-06 ***
6:NrZanore             1.4534e-01  7.0084e-02  2.0737e+00 0.0381054 *
```

```
6:NrBashketingellore  -1.8524e-01  6.4088e-02 -2.8903e+00 0.0038484 **
6:NrFjali             -3.3842e+00  1.8081e-01 -1.8717e+01 < 2.2e-16 ***
6:NrShenjapikesimi     2.0263e+00  6.7536e-02  3.0004e+01 < 2.2e-16 ***
6:NrQe                -3.3834e+00  1.7838e-01 -1.8967e+01 < 2.2e-16 ***
6:NrSE                 9.1842e+00  7.7759e-02  1.1811e+02 < 2.2e-16 ***
7:(Intercept)          1.1968e+01  2.5307e-03  4.7292e+03 < 2.2e-16 ***
7:NrFjale              3.8864e-01  7.0155e-02  5.5397e+00 3.030e-08 ***
7:NrGerma             -5.7998e-02  9.0253e-03 -6.4262e+00 1.308e-10 ***
7:NrZanore            -3.5409e-03  1.4933e-02 -2.3710e-01 0.8125639
7:NrBashketingellore  -5.4457e-02  1.6842e-02 -3.2334e+00 0.0012232 **
7:NrFjali             -3.0379e+00  8.5259e-02 -3.5632e+01 < 2.2e-16 ***
7:NrShenjapikesimi     2.1550e+00  7.0058e-02  3.0760e+01 < 2.2e-16 ***
7:NrQe                -3.6915e+00  1.0390e-01 -3.5531e+01 < 2.2e-16 ***
7:NrSE                 4.1815e+00  1.3634e-01  3.0669e+01 < 2.2e-16 ***
8:(Intercept)          1.3071e+01  3.0862e-03  4.2351e+03 < 2.2e-16 ***
8:NrFjale              3.6447e-01  7.1900e-02  5.0692e+00 3.996e-07 ***
8:NrGerma             -5.5973e-02  9.1880e-03 -6.0920e+00 1.115e-09 ***
8:NrZanore            -3.3782e-03  1.4944e-02 -2.2610e-01 0.8211525
8:NrBashketingellore  -5.2595e-02  1.6949e-02 -3.1032e+00 0.0019146 **
8:NrFjali             -2.9187e+00  8.7131e-02 -3.3498e+01 < 2.2e-16 ***
8:NrShenjapikesimi     2.1737e+00  6.9486e-02  3.1283e+01 < 2.2e-16 ***
8:NrQe                -3.7562e+00  1.0577e-01 -3.5511e+01 < 2.2e-16 ***
8:NrSE                 4.2676e+00  1.4827e-01  2.8783e+01 < 2.2e-16 ***
9:(Intercept)          1.1848e+01  3.1104e-03  3.8090e+03 < 2.2e-16 ***
9:NrFjale              5.0834e-01  6.5431e-02  7.7691e+00 7.904e-15 ***
9:NrGerma             -7.3801e-02  8.7847e-03 -8.4010e+00 < 2.2e-16 ***
9:NrZanore            -9.0475e-03  1.5764e-02 -5.7390e-01 0.5660047
9:NrBashketingellore  -6.4753e-02  1.7406e-02 -3.7202e+00 0.0001991 ***
9:NrFjali             -3.3544e+00  8.7410e-02 -3.8376e+01 < 2.2e-16 ***
9:NrShenjapikesimi     2.2184e+00  6.9434e-02  3.1950e+01 < 2.2e-16 ***
9:NrQe                -3.4707e+00  8.4822e-02 -4.0917e+01 < 2.2e-16 ***
9:NrSE                 4.0419e+00  1.2259e-01  3.2971e+01 < 2.2e-16 ***
10:(Intercept)        -2.8383e+00  5.7824e-03 -4.9085e+02 < 2.2e-16 ***
10:NrFjale             4.8516e-01  6.5117e-02  7.4505e+00 9.298e-14 ***
10:NrGerma            -6.8243e-02  8.6686e-03 -7.8724e+00 3.479e-15 ***
10:NrZanore           -9.9597e-04  1.6622e-02 -5.9900e-02 0.9522198
10:NrBashketingellore -6.7247e-02  1.8067e-02 -3.7221e+00 0.0001976 ***
10:NrFjali            -3.3180e+00  7.3601e-02 -4.5082e+01 < 2.2e-16 ***
10:NrShenjapikesimi    2.2639e+00  6.5637e-02  3.4491e+01 < 2.2e-16 ***
10:NrQe               -3.7033e+00  1.1198e-01 -3.3072e+01 < 2.2e-16 ***
10:NrSE                3.8489e+00  1.6616e-01  2.3164e+01 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
> Pred1<-predict(MLARTSH1,DATAARTSH)
> PredMult1<-matrix(Pred1,nrow=10,byrow=TRUE)
> PredMult1
      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
 [1,] "1"  "1"  "1"  "1"  "1"  "1"  "1"  "1"  "1"  "1"
 [2,] "2"  "2"  "2"  "2"  "2"  "2"  "2"  "2"  "2"  "2"
 [3,] "3"  "3"  "3"  "3"  "3"  "3"  "3"  "3"  "3"  "3"
 [4,] "4"  "4"  "4"  "4"  "4"  "4"  "4"  "4"  "4"  "4"
 [5,] "5"  "9"  "5"  "5"  "5"  "5"  "5"  "5"  "5"  "5"
 [6,] "6"  "6"  "6"  "6"  "6"  "6"  "6"  "6"  "6"  "6"
 [7,] "9"  "8"  "7"  "7"  "7"  "7"  "7"  "6"  "8"  "7"
```

```
 [8,] "8"  "8"  "8"  "8"  "7"  "7"  "7"  "7"  "8"  "8"
 [9,] "10" "10" "8"  "9"  "9"  "9"  "8"  "10" "10" "9"
[10,] "10" "5"  "9"  "7"  "10" "10" "9"  "10" "10" "10"
```

Rezultatet e matricës PredMult1 janë përmbledhur në tabelën 3.10. Saktësia e klasifikimit të teksteve ne modelin MLARTSH1 arrin në 81 %. Vërehet qe kemi një rritje shumë të madhe të përqindes se klasifikimit të sakte krahasuar me modelin e ndertar në Salillari, Prifti(2016) në të cilin saktësia rezultoi 59%.

AUTORI	Klasifikimi i sakte në %
1	100
2	100
3	100
4	100
5	90
6	100
7	60
8	60
9	40
10	60
Mesatarja	81

TABELA 3. 10: REZULTATET E KLASIFIKIMIT NE MODELIN MLARTSH1

Në këtë model vërehet që variabli NrZanore nuk është statistikisht i rëndësishëm.

Kështu që më poshtë po shqyrtohet një model ku e përjashtohet variabli NrZanore.

```
> MLARTSH2<-multinom(AUTORI~NrFjale+NrGerma+NrBashketingellore+NrFjali+NrShenjapikesimi+NrQe+NrSE,data=DATAARTSH)
# weights:  90 (72 variable)
initial  value 230.258509
iter  10 value 200.772017
iter  20 value 132.813304
iter  30 value 120.200000
iter  40 value 116.317684
iter  50 value 97.440042
iter  60 value 80.980268
iter  70 value 58.510256
iter  80 value 51.591495
iter  90 value 46.912278
iter 100 value 45.006014
```

```
final  value 45.006014
stopped after 100 iterations
> summary(MLARTSH2)
Call:
multinom(formula = AUTORI ~ NrFjale + NrGerma + NrBashketingellore +
    NrFjali + NrShenjapikesimi + NrQe + NrSE, data = DATAARTSH)

Coefficients:
   (Intercept)     NrFjale    NrGerma NrBashketingellore   NrFjali NrShenjapikesimi
 NrQe      NrSE
2  -151.309000 0.006443215 -0.1664758         0.23938971 -3.703605         3.601251  -7.3
02425 12.527224
3    99.378200 1.196319098 -0.3288573         0.13799007 -9.838238         3.062541  -5.9
11818  8.931713
4   -28.605330 1.103624111  0.1063639        -0.53712366 -3.825509         2.681153 -10.8
24412  3.576496
5     6.362407 0.798450013 -0.1512271        -0.03861836 -5.138120         3.172183  -5.7
20203  7.043140
6   -60.804961 0.627662451  0.2078171        -0.60164473 -4.893219         3.225033  -5.3
17864 12.409245
7     8.825941 0.666956989 -0.1174538        -0.05305370 -4.993146         3.209491  -5.7
90465  6.951484
8     9.835919 0.640766783 -0.1156974        -0.05020677 -4.876020         3.234627  -5.8
59130  7.064575
9    10.306933 0.767981623 -0.1369017        -0.05561282 -5.265828         3.270458  -5.5
92622  6.844890
10   -2.957843 0.744423571 -0.1243520        -0.06441758 -5.207464         3.302463  -5.8
29779  6.687153

Std. Errors:
    (Intercept)     NrFjale     NrGerma NrBashketingellore     NrFjali NrShenjapikesimi
       NrQe
2  1.109316e-03 0.095364526 0.038865736        0.052706469 0.130304397      0.137587733 0
.1572786607
3  1.749084e-03 0.144864618 0.106681935        0.166631279 0.037127190      0.108593812 0
.1467495716
4  7.345147e-06 0.009740799 0.006744907        0.003710056 0.001693379      0.003452557 0
.0003796908
5  5.796428e-03 0.083488163 0.039258050        0.078684138 0.066566331      0.067466192 0
.0918225196
6  9.475813e-03 0.074947992 0.082735820        0.153401954 0.113526976      0.127440863 0
.1230214622
7  2.246707e-03 0.083836354 0.039089951        0.078567655 0.069233791      0.066010699 0
.0993219089
8  2.248270e-03 0.085367219 0.039064553        0.078554643 0.072584881      0.066083138 0
.1020405866
9  6.376061e-03 0.081421872 0.039216907        0.079344746 0.079387830      0.064212078 0
.0865230494
10 5.739649e-03 0.081408179 0.038937029        0.079141102 0.062716572      0.060651696 0
.1011426158
           NrSE
2  0.0212715608
3  0.0917637097
4  0.0002395968
5  0.1411920899
```

```
6  0.0843963548
7  0.1335266506
8  0.1453161277
9  0.1193579062
10 0.1524734996

Residual Deviance: 90.01203
AIC: 234.012
> coeftest(MLARTSH2)

z test of coefficients:

                        Estimate  Std. Error     z value  Pr(>|z|)
2:(Intercept)         -1.5131e+02  1.1093e-03 -1.3640e+05 < 2.2e-16 ***
2:NrFjale              6.4432e-03  9.5365e-02  6.7600e-02 0.9461327
2:NrGerma             -1.6648e-01  3.8866e-02 -4.2834e+00 1.841e-05 ***
2:NrBashketingellore   2.3939e-01  5.2706e-02  4.5419e+00 5.574e-06 ***
2:NrFjali             -3.7036e+00  1.3030e-01 -2.8423e+01 < 2.2e-16 ***
2:NrShenjapikesimi     3.6013e+00  1.3759e-01  2.6174e+01 < 2.2e-16 ***
2:NrQe                -7.3024e+00  1.5728e-01 -4.6430e+01 < 2.2e-16 ***
2:NrSE                 1.2527e+01  2.1272e-02  5.8892e+02 < 2.2e-16 ***
3:(Intercept)          9.9378e+01  1.7491e-03  5.6817e+04 < 2.2e-16 ***
3:NrFjale              1.1963e+00  1.4486e-01  8.2582e+00 < 2.2e-16 ***
3:NrGerma             -3.2886e-01  1.0668e-01 -3.0826e+00 0.0020520 **
3:NrBashketingellore   1.3799e-01  1.6663e-01  8.2810e-01 0.4076047
3:NrFjali             -9.8382e+00  3.7127e-02 -2.6499e+02 < 2.2e-16 ***
3:NrShenjapikesimi     3.0625e+00  1.0859e-01  2.8202e+01 < 2.2e-16 ***
3:NrQe                -5.9118e+00  1.4675e-01 -4.0285e+01 < 2.2e-16 ***
3:NrSE                 8.9317e+00  9.1764e-02  9.7334e+01 < 2.2e-16 ***
4:(Intercept)         -2.8605e+01  7.3451e-06 -3.8945e+06 < 2.2e-16 ***
4:NrFjale              1.1036e+00  9.7408e-03  1.1330e+02 < 2.2e-16 ***
4:NrGerma              1.0636e-01  6.7449e-03  1.5770e+01 < 2.2e-16 ***
4:NrBashketingellore  -5.3712e-01  3.7101e-03 -1.4478e+02 < 2.2e-16 ***
4:NrFjali             -3.8255e+00  1.6934e-03 -2.2591e+03 < 2.2e-16 ***
4:NrShenjapikesimi     2.6812e+00  3.4526e-03  7.7657e+02 < 2.2e-16 ***
4:NrQe                -1.0824e+01  3.7969e-04 -2.8508e+04 < 2.2e-16 ***
4:NrSE                 3.5765e+00  2.3960e-04  1.4927e+04 < 2.2e-16 ***
5:(Intercept)          6.3624e+00  5.7964e-03  1.0976e+03 < 2.2e-16 ***
5:NrFjale              7.9845e-01  8.3488e-02  9.5636e+00 < 2.2e-16 ***
5:NrGerma             -1.5123e-01  3.9258e-02 -3.8521e+00 0.0001171 ***
5:NrBashketingellore  -3.8618e-02  7.8684e-02 -4.9080e-01 0.6235662
5:NrFjali             -5.1381e+00  6.6566e-02 -7.7188e+01 < 2.2e-16 ***
5:NrShenjapikesimi     3.1722e+00  6.7466e-02  4.7019e+01 < 2.2e-16 ***
5:NrQe                -5.7202e+00  9.1823e-02 -6.2296e+01 < 2.2e-16 ***
5:NrSE                 7.0431e+00  1.4119e-01  4.9883e+01 < 2.2e-16 ***
6:(Intercept)         -6.0805e+01  9.4758e-03 -6.4169e+03 < 2.2e-16 ***
6:NrFjale              6.2766e-01  7.4948e-02  8.3746e+00 < 2.2e-16 ***
6:NrGerma              2.0782e-01  8.2736e-02  2.5118e+00 0.0120112 *
6:NrBashketingellore  -6.0164e-01  1.5340e-01 -3.9220e+00 8.781e-05 ***
6:NrFjali             -4.8932e+00  1.1353e-01 -4.3102e+01 < 2.2e-16 ***
6:NrShenjapikesimi     3.2250e+00  1.2744e-01  2.5306e+01 < 2.2e-16 ***
6:NrQe                -5.3179e+00  1.2302e-01 -4.3227e+01 < 2.2e-16 ***
6:NrSE                 1.2409e+01  8.4396e-02  1.4704e+02 < 2.2e-16 ***
7:(Intercept)          8.8259e+00  2.2467e-03  3.9284e+03 < 2.2e-16 ***
7:NrFjale              6.6696e-01  8.3836e-02  7.9555e+00 1.785e-15 ***
```

```
7:NrGerma             -1.1745e-01  3.9090e-02 -3.0047e+00 0.0026584 **
7:NrBashketingellore  -5.3054e-02  7.8568e-02 -6.7530e-01 0.4995097
7:NrFjali             -4.9931e+00  6.9234e-02 -7.2120e+01 < 2.2e-16 ***
7:NrShenjapikesimi     3.2095e+00  6.6011e-02  4.8621e+01 < 2.2e-16 ***
7:NrQe                -5.7905e+00  9.9322e-02 -5.8300e+01 < 2.2e-16 ***
7:NrSE                 6.9515e+00  1.3353e-01  5.2061e+01 < 2.2e-16 ***
8:(Intercept)          9.8359e+00  2.2483e-03  4.3749e+03 < 2.2e-16 ***
8:NrFjale              6.4077e-01  8.5367e-02  7.5060e+00 6.096e-14 ***
8:NrGerma             -1.1570e-01  3.9065e-02 -2.9617e+00 0.0030595 **
8:NrBashketingellore  -5.0207e-02  7.8555e-02 -6.3910e-01 0.5227372
8:NrFjali             -4.8760e+00  7.2585e-02 -6.7177e+01 < 2.2e-16 ***
8:NrShenjapikesimi     3.2346e+00  6.6083e-02  4.8948e+01 < 2.2e-16 ***
8:NrQe                -5.8591e+00  1.0204e-01 -5.7420e+01 < 2.2e-16 ***
8:NrSE                 7.0646e+00  1.4532e-01  4.8615e+01 < 2.2e-16 ***
9:(Intercept)          1.0307e+01  6.3761e-03  1.6165e+03 < 2.2e-16 ***
9:NrFjale              7.6798e-01  8.1422e-02  9.4321e+00 < 2.2e-16 ***
9:NrGerma             -1.3690e-01  3.9217e-02 -3.4909e+00 0.0004814 ***
9:NrBashketingellore  -5.5613e-02  7.9345e-02 -7.0090e-01 0.4833648
9:NrFjali             -5.2658e+00  7.9388e-02 -6.6330e+01 < 2.2e-16 ***
9:NrShenjapikesimi     3.2705e+00  6.4212e-02  5.0932e+01 < 2.2e-16 ***
9:NrQe                -5.5926e+00  8.6523e-02 -6.4637e+01 < 2.2e-16 ***
9:NrSE                 6.8449e+00  1.1936e-01  5.7348e+01 < 2.2e-16 ***
10:(Intercept)        -2.9578e+00  5.7396e-03 -5.1534e+02 < 2.2e-16 ***
10:NrFjale             7.4442e-01  8.1408e-02  9.1443e+00 < 2.2e-16 ***
10:NrGerma            -1.2435e-01  3.8937e-02 -3.1937e+00 0.0014048 **
10:NrBashketingellore -6.4418e-02  7.9141e-02 -8.1400e-01 0.4156687
10:NrFjali            -5.2075e+00  6.2717e-02 -8.3032e+01 < 2.2e-16 ***
10:NrShenjapikesimi    3.3025e+00  6.0652e-02  5.4450e+01 < 2.2e-16 ***
10:NrQe               -5.8298e+00  1.0114e-01 -5.7639e+01 < 2.2e-16 ***
10:NrSE                6.6872e+00  1.5247e-01  4.3858e+01 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
> Pred2<-predict(MLARTSH2,DATAARTSH)
> PredMult2<-matrix(Pred2,nrow=10,byrow=TRUE)
> PredMult2
```

```
      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
 [1,] "1"  "1"  "1"  "1"  "1"  "1"  "1"  "1"  "1"  "1"
 [2,] "2"  "2"  "2"  "2"  "2"  "2"  "2"  "2"  "2"  "2"
 [3,] "3"  "3"  "3"  "3"  "3"  "3"  "3"  "3"  "3"  "3"
 [4,] "4"  "4"  "4"  "4"  "4"  "4"  "4"  "4"  "4"  "4"
 [5,] "5"  "9"  "5"  "5"  "5"  "5"  "5"  "5"  "5"  "5"
 [6,] "6"  "6"  "6"  "6"  "6"  "6"  "6"  "6"  "6"  "6"
 [7,] "9"  "8"  "7"  "7"  "7"  "7"  "7"  "6"  "8"  "7"
 [8,] "8"  "8"  "8"  "8"  "7"  "7"  "7"  "7"  "8"  "8"
 [9,] "10" "10" "8"  "9"  "9"  "9"  "8"  "10" "10" "9"
[10,] "10" "9"  "10" "7"  "10" "10" "9"  "10" "10" "10"
```

Rezultatet e matricës PredMult2 janë përmbledhur në tabelën 3.11. Saktësia e klasifikimit të teksteve në modelin MLARTSH2 mbetet e njëjtë 82 %. Vërehet që parametrat janë thuajse te gjithe statistikisht te rendesishem me α=0.001.

AUTORI	Klasifikimi i sakte në %
1	100
2	100
3	100
4	100
5	90
6	100
7	60
8	60
9	40
10	70
Mesatarja	82

TABELA 3. 11: REZULTATET E KLASIFIKIMIT NE MODELIN MLARTSH2

Duke patur parasysh modelin më të mirë të ndërtar në Salillari,Prifti(2016) ndërtohet modeli kur përjashtohet vetëm numri i germave si më poshtë:

```
> MLARTSH3<-multinom(AUTORI~NrFjale+NrZanore+NrBashketingellore+NrFjali+NrShenjapikesimi+
NrQe+NrSE,data=DATAARTSH)
# weights:  90 (72 variable)
initial  value 230.258509
iter  10 value 209.352939
iter  20 value 150.266378
iter  30 value 133.976412
iter  40 value 127.585746
iter  50 value 103.909045
iter  60 value 91.429748
iter  70 value 62.898556
iter  80 value 52.369584
iter  90 value 48.030088
iter 100 value 43.735656
final  value 43.735656
stopped after 100 iterations
> summary(MLARTSH3)
Call:
multinom(formula = AUTORI ~ NrFjale + NrZanore + NrBashketingellore +
    NrFjali + NrShenjapikesimi + NrQe + NrSE, data = DATAARTSH)

Coefficients:
   (Intercept)    NrFjale     NrZanore NrBashketingellore    NrFjali NrShenjapikesimi
   NrQe       NrSE
2  -246.539446 -0.2176779 -0.200554275          0.1646150  -4.457938         4.520538  -9
.771280 18.1455544
```

```
3   211.060818  1.2777917 -0.423801669         -0.1408825 -14.162233         3.144636  -6
.073721 10.1681137
4   -55.427241  1.6488164 -0.003336117         -0.5485685  -5.329279         4.013703 -13
.736042  0.9790706
5    18.872275  0.9262292 -0.183774531         -0.2188191  -6.469128         3.897237  -7
.097195  9.1549068
6  -110.304911  0.6263550  0.339039736         -0.4838202  -6.670250         3.513709  -6
.507293 20.8008876
7    21.396666  0.7858241 -0.148824049         -0.1977198  -6.310843         3.946891  -7
.169516  9.0012106
8    21.799241  0.7602201 -0.146860347         -0.1935551  -6.197923         3.980740  -7
.244789  9.1163022
9    22.100924  0.8873845 -0.167281107         -0.2199156  -6.582244         3.992868  -6
.957326  8.9653394
10    8.075254  0.8708403 -0.156099977         -0.2177771  -6.559051         4.055594  -7
.235432  8.7101746

Std. Errors:
    (Intercept)    NrFjale    NrZanore NrBashketingellore      NrFjali NrShenjapikesimi
      NrQe
2  2.473988e-03 0.23749020 0.042403624         0.05342467 0.3447735493     0.3268637582 0
.0995774979
3  2.269996e-03 0.24229127 0.134607117         0.09743365 0.0910079702     0.1442860805 0
.2087675951
4  8.053323e-06 0.00237175 0.004612158         0.00751920 0.0004600652     0.0008281525 0
.0001485026
5  5.704979e-03 0.12320651 0.109340803         0.07960166 0.0886100299     0.1104848224 0
.1259507571
6  1.232240e-02 0.15872806 0.168161672         0.15311385 0.2090794416     0.2549635003 0
.3147334182
7  3.114114e-03 0.12215980 0.109292036         0.07930970 0.0882182789     0.1073648190 0
.1246360354
8  2.923105e-03 0.12305042 0.109308743         0.07941050 0.0893429212     0.1068456364 0
.1261807544
9  1.004153e-02 0.12188827 0.108952897         0.07971151 0.0973333121     0.1077008712 0
.1198242481
10 6.046792e-03 0.12075667 0.109568088         0.07968159 0.0840296585     0.1056668203 0
.1367854744
           NrSE
2  4.703141e-02
3  9.551303e-02
4  7.546566e-05
5  1.494726e-01
6  7.821502e-02
7  1.392211e-01
8  1.517302e-01
9  1.198609e-01
10 1.639609e-01

Residual Deviance: 87.47131
AIC: 231.4713
> coeftest(MLARTSH3)

z test of coefficients:
```

```
                      Estimate  Std. Error     z value  Pr(>|z|)
2:(Intercept)        -2.4654e+02  2.4740e-03 -9.9653e+04 < 2.2e-16 ***
2:NrFjale            -2.1768e-01  2.3749e-01 -9.1660e-01  0.359365
2:NrZanore           -2.0055e-01  4.2404e-02 -4.7296e+00 2.249e-06 ***
2:NrBashketingellore  1.6462e-01  5.3425e-02  3.0813e+00  0.002061 **
2:NrFjali            -4.4579e+00  3.4477e-01 -1.2930e+01 < 2.2e-16 ***
2:NrShenjapikesimi    4.5205e+00  3.2686e-01  1.3830e+01 < 2.2e-16 ***
2:NrQe               -9.7713e+00  9.9577e-02 -9.8127e+01 < 2.2e-16 ***
2:NrSE                1.8146e+01  4.7031e-02  3.8582e+02 < 2.2e-16 ***
3:(Intercept)         2.1106e+02  2.2700e-03  9.2979e+04 < 2.2e-16 ***
3:NrFjale             1.2778e+00  2.4229e-01  5.2738e+00 1.336e-07 ***
3:NrZanore           -4.2380e-01  1.3461e-01 -3.1484e+00  0.001641 **
3:NrBashketingellore -1.4088e-01  9.7434e-02 -1.4459e+00  0.148196
3:NrFjali            -1.4162e+01  9.1008e-02 -1.5562e+02 < 2.2e-16 ***
3:NrShenjapikesimi    3.1446e+00  1.4429e-01  2.1794e+01 < 2.2e-16 ***
3:NrQe               -6.0737e+00  2.0877e-01 -2.9093e+01 < 2.2e-16 ***
3:NrSE                1.0168e+01  9.5513e-02  1.0646e+02 < 2.2e-16 ***
4:(Intercept)        -5.5427e+01  8.0533e-06 -6.8825e+06 < 2.2e-16 ***
4:NrFjale             1.6488e+00  2.3717e-03  6.9519e+02 < 2.2e-16 ***
4:NrZanore           -3.3361e-03  4.6122e-03 -7.2330e-01  0.469477
4:NrBashketingellore -5.4857e-01  7.5192e-03 -7.2956e+01 < 2.2e-16 ***
4:NrFjali            -5.3293e+00  4.6007e-04 -1.1584e+04 < 2.2e-16 ***
4:NrShenjapikesimi    4.0137e+00  8.2815e-04  4.8466e+03 < 2.2e-16 ***
4:NrQe               -1.3736e+01  1.4850e-04 -9.2497e+04 < 2.2e-16 ***
4:NrSE                9.7907e-01  7.5466e-05  1.2974e+04 < 2.2e-16 ***
5:(Intercept)         1.8872e+01  5.7050e-03  3.3080e+03 < 2.2e-16 ***
5:NrFjale             9.2623e-01  1.2321e-01  7.5177e+00 5.575e-14 ***
5:NrZanore           -1.8377e-01  1.0934e-01 -1.6807e+00  0.092812 .
5:NrBashketingellore -2.1882e-01  7.9602e-02 -2.7489e+00  0.005979 **
5:NrFjali            -6.4691e+00  8.8610e-02 -7.3007e+01 < 2.2e-16 ***
5:NrShenjapikesimi    3.8972e+00  1.1048e-01  3.5274e+01 < 2.2e-16 ***
5:NrQe               -7.0972e+00  1.2595e-01 -5.6349e+01 < 2.2e-16 ***
5:NrSE                9.1549e+00  1.4947e-01  6.1248e+01 < 2.2e-16 ***
6:(Intercept)        -1.1030e+02  1.2322e-02 -8.9516e+03 < 2.2e-16 ***
6:NrFjale             6.2636e-01  1.5873e-01  3.9461e+00 7.944e-05 ***
6:NrZanore            3.3904e-01  1.6816e-01  2.0162e+00  0.043784 *
6:NrBashketingellore -4.8382e-01  1.5311e-01 -3.1599e+00  0.001578 **
6:NrFjali            -6.6702e+00  2.0908e-01 -3.1903e+01 < 2.2e-16 ***
6:NrShenjapikesimi    3.5137e+00  2.5496e-01  1.3781e+01 < 2.2e-16 ***
6:NrQe               -6.5073e+00  3.1473e-01 -2.0676e+01 < 2.2e-16 ***
6:NrSE                2.0801e+01  7.8215e-02  2.6594e+02 < 2.2e-16 ***
7:(Intercept)         2.1397e+01  3.1141e-03  6.8709e+03 < 2.2e-16 ***
7:NrFjale             7.8582e-01  1.2216e-01  6.4328e+00 1.253e-10 ***
7:NrZanore           -1.4882e-01  1.0929e-01 -1.3617e+00  0.173289
7:NrBashketingellore -1.9772e-01  7.9310e-02 -2.4930e+00  0.012667 *
7:NrFjali            -6.3108e+00  8.8218e-02 -7.1537e+01 < 2.2e-16 ***
7:NrShenjapikesimi    3.9469e+00  1.0736e-01  3.6761e+01 < 2.2e-16 ***
7:NrQe               -7.1695e+00  1.2464e-01 -5.7524e+01 < 2.2e-16 ***
7:NrSE                9.0012e+00  1.3922e-01  6.4654e+01 < 2.2e-16 ***
8:(Intercept)         2.1799e+01  2.9231e-03  7.4576e+03 < 2.2e-16 ***
8:NrFjale             7.6022e-01  1.2305e-01  6.1781e+00 6.487e-10 ***
8:NrZanore           -1.4686e-01  1.0931e-01 -1.3435e+00  0.179098
8:NrBashketingellore -1.9356e-01  7.9410e-02 -2.4374e+00  0.014793 *
8:NrFjali            -6.1979e+00  8.9343e-02 -6.9372e+01 < 2.2e-16 ***
8:NrShenjapikesimi    3.9807e+00  1.0685e-01  3.7257e+01 < 2.2e-16 ***
```

```
8:NrQe                -7.2448e+00  1.2618e-01 -5.7416e+01 < 2.2e-16 ***
8:NrSE                 9.1163e+00  1.5173e-01  6.0082e+01 < 2.2e-16 ***
9:(Intercept)          2.2101e+01  1.0042e-02  2.2010e+03 < 2.2e-16 ***
9:NrFjale              8.8738e-01  1.2189e-01  7.2803e+00 3.331e-13 ***
9:NrZanore            -1.6728e-01  1.0895e-01 -1.5354e+00  0.124697
9:NrBashketingellore  -2.1992e-01  7.9712e-02 -2.7589e+00  0.005800 **
9:NrFjali             -6.5822e+00  9.7333e-02 -6.7626e+01 < 2.2e-16 ***
9:NrShenjapikesimi     3.9929e+00  1.0770e-01  3.7074e+01 < 2.2e-16 ***
9:NrQe                -6.9573e+00  1.1982e-01 -5.8063e+01 < 2.2e-16 ***
9:NrSE                 8.9653e+00  1.1986e-01  7.4798e+01 < 2.2e-16 ***
10:(Intercept)         8.0753e+00  6.0468e-03  1.3355e+03 < 2.2e-16 ***
10:NrFjale             8.7084e-01  1.2076e-01  7.2115e+00 5.533e-13 ***
10:NrZanore           -1.5610e-01  1.0957e-01 -1.4247e+00  0.154248
10:NrBashketingellore -2.1778e-01  7.9682e-02 -2.7331e+00  0.006274 **
10:NrFjali            -6.5591e+00  8.4030e-02 -7.8056e+01 < 2.2e-16 ***
10:NrShenjapikesimi    4.0556e+00  1.0567e-01  3.8381e+01 < 2.2e-16 ***
10:NrQe               -7.2354e+00  1.3679e-01 -5.2896e+01 < 2.2e-16 ***
10:NrSE                8.7102e+00  1.6396e-01  5.3123e+01 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
> Pred3<- predict(MLARTSH3,DATAARTSH)
> PredMult3<-matrix(Pred3,nrow=10,byrow=TRUE)
> PredMult3
      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
 [1,] "1"  "1"  "1"  "1"  "1"  "1"  "1"  "1"  "1"  "1"
 [2,] "2"  "2"  "2"  "2"  "2"  "2"  "2"  "2"  "2"  "2"
 [3,] "3"  "3"  "3"  "3"  "3"  "3"  "3"  "3"  "3"  "3"
 [4,] "4"  "4"  "4"  "4"  "4"  "4"  "4"  "4"  "4"  "4"
 [5,] "5"  "9"  "5"  "5"  "5"  "5"  "5"  "5"  "9"  "5"
 [6,] "6"  "6"  "6"  "6"  "6"  "6"  "6"  "6"  "6"  "6"
 [7,] "9"  "8"  "7"  "7"  "7"  "7"  "7"  "7"  "8"  "7"
 [8,] "8"  "8"  "8"  "8"  "7"  "7"  "7"  "7"  "8"  "8"
 [9,] "10" "10" "7"  "9"  "9"  "9"  "9"  "10" "10" "9"
[10,] "10" "9"  "9"  "7"  "10" "10" "9"  "10" "10" "10"
```

Nga ky model vërehet se probabiliteti i klasifikimit saktë të teksteve nuk ndryshon mbetet përsëri 0.82 , veçse variabli NrZanore del statistikisht i parëndësishëm për $\alpha=0.05$.

Në modelin e mëposhtëm përjashtohet edhe variabli NrZanore.

```
>MLARTSH4<-multinom(AUTORI~NrFjale+NrBashketingellore+NrFjali+NrShenjapikesimi+NrQe+NrSE,
data=DATAARTSH)
# weights:  80 (63 variable)
initial  value 230.258509
iter  10 value 199.738864
iter  20 value 138.967923
iter  30 value 129.631473
iter  40 value 107.381679
```

```
iter  50 value 99.285524
iter  60 value 71.412391
iter  70 value 65.611215
iter  80 value 64.047429
iter  90 value 62.387458
iter 100 value 61.678669
final  value 61.678669
stopped after 100 iterations
> summary(MLARTSH4)
Call:
multinom(formula = AUTORI ~ NrFjale + NrBashketingellore + NrFjali +
    NrShenjapikesimi + NrQe + NrSE, data = DATAARTSH)

Coefficients:
   (Intercept)        NrFjale  NrBashketingellore    NrFjali  NrShenjapikesimi        NrQe
NrSE
2    -42.212139 -0.007063334          -0.01607520 -1.756692           1.664482 -2.867288
3.743457
3     66.039247   0.334050634          -0.13022062 -5.304568           1.764641 -2.279407
3.354052
4    -12.256187   0.477264783          -0.16290348 -1.732552           1.461488 -4.838007 -
1.082430
5     11.789626   0.169112143          -0.07956799 -2.217840           1.672180 -2.105540
1.906169
6    -28.070867   0.068974368          -0.03062857 -2.853470           1.151119 -2.110684
7.038639
7     12.976661   0.111624366          -0.06041191 -2.222906           1.700710 -2.209006
1.876718
8     13.150658   0.095811915          -0.05703435 -2.109962           1.710251 -2.279875
1.955574
9     15.087285   0.158699499          -0.07919570 -2.378971           1.765784 -1.993674
1.823863
10     5.537583   0.154727261          -0.07472105 -2.360542           1.783931 -2.172552
1.832205


Std. Errors:
    (Intercept)     NrFjale NrBashketingellore      NrFjali NrShenjapikesimi         NrQe
NrSE
2  1.265630e-02 0.038504694        0.011490197 0.0799313695     0.1091761347 1.165554e-01
2.413303e-01
3  4.093331e-03 0.211319917        0.079627790 0.1563448612     0.2957321077 2.423760e-01
1.725831e-01
4  2.544686e-06 0.003511888        0.009433748 0.0001312502     0.0003593038 5.384853e-05
3.579452e-05
5  1.936702e-02 0.038591877        0.011784999 0.1094796023     0.0817221023 1.391395e-01
1.616403e-01
6  5.745132e-02 0.068091744        0.020321346 0.1744879024     0.1138563039 2.243338e-01
3.969701e-01
7  4.528213e-03 0.035987653        0.010633920 0.1082744074     0.0675307148 1.250223e-01
1.644004e-01
8  3.942599e-03 0.035076862        0.010352029 0.1060540002     0.0696005798 1.215656e-01
1.674438e-01
```

```
9  2.291895e-02 0.039095004         0.011985773 0.1227683095     0.0827859896 1.404609e-01
1.523548e-01
10 1.204655e-02 0.038019346         0.011473991 0.1135241763     0.0748416242 1.395955e-01
1.853254e-01


Residual Deviance: 123.3573
AIC: 249.3573
> coeftest(MLARTSH4)
z test of coefficients:
                         Estimate  Std. Error     z value  Pr(>|z|)
2:(Intercept)         -4.2212e+01  1.2656e-02 -3.3353e+03 < 2.2e-16 ***
2:NrFjale             -7.0633e-03  3.8505e-02 -1.8340e-01  0.854452
2:NrBashketingellore  -1.6075e-02  1.1490e-02 -1.3990e+00  0.161802
2:NrFjali             -1.7567e+00  7.9931e-02 -2.1977e+01 < 2.2e-16 ***
2:NrShenjapikesimi     1.6645e+00  1.0918e-01  1.5246e+01 < 2.2e-16 ***
2:NrQe                -2.8673e+00  1.1656e-01 -2.4600e+01 < 2.2e-16 ***
2:NrSE                 3.7435e+00  2.4133e-01  1.5512e+01 < 2.2e-16 ***
3:(Intercept)          6.6039e+01  4.0933e-03  1.6133e+04 < 2.2e-16 ***
3:NrFjale              3.3405e-01  2.1132e-01  1.5808e+00  0.113928
3:NrBashketingellore  -1.3022e-01  7.9628e-02 -1.6354e+00  0.101972
3:NrFjali             -5.3046e+00  1.5634e-01 -3.3929e+01 < 2.2e-16 ***
3:NrShenjapikesimi     1.7646e+00  2.9573e-01  5.9670e+00 2.416e-09 ***
3:NrQe                -2.2794e+00  2.4238e-01 -9.4044e+00 < 2.2e-16 ***
3:NrSE                 3.3541e+00  1.7258e-01  1.9434e+01 < 2.2e-16 ***
4:(Intercept)         -1.2256e+01  2.5447e-06 -4.8164e+06 < 2.2e-16 ***
4:NrFjale              4.7726e-01  3.5119e-03  1.3590e+02 < 2.2e-16 ***
4:NrBashketingellore  -1.6290e-01  9.4337e-03 -1.7268e+01 < 2.2e-16 ***
4:NrFjali             -1.7326e+00  1.3125e-04 -1.3200e+04 < 2.2e-16 ***
4:NrShenjapikesimi     1.4615e+00  3.5930e-04  4.0676e+03 < 2.2e-16 ***
4:NrQe                -4.8380e+00  5.3848e-05 -8.9845e+04 < 2.2e-16 ***
4:NrSE                -1.0824e+00  3.5794e-05 -3.0240e+04 < 2.2e-16 ***
5:(Intercept)          1.1790e+01  1.9367e-02  6.0875e+02 < 2.2e-16 ***
5:NrFjale              1.6911e-01  3.8592e-02  4.3821e+00 1.176e-05 ***
5:NrBashketingellore  -7.9568e-02  1.1785e-02 -6.7516e+00 1.462e-11 ***
5:NrFjali             -2.2178e+00  1.0948e-01 -2.0258e+01 < 2.2e-16 ***
5:NrShenjapikesimi     1.6722e+00  8.1722e-02  2.0462e+01 < 2.2e-16 ***
5:NrQe                -2.1055e+00  1.3914e-01 -1.5133e+01 < 2.2e-16 ***
5:NrSE                 1.9062e+00  1.6164e-01  1.1793e+01 < 2.2e-16 ***
6:(Intercept)         -2.8071e+01  5.7451e-02 -4.8860e+02 < 2.2e-16 ***
6:NrFjale              6.8974e-02  6.8092e-02  1.0130e+00  0.311078
6:NrBashketingellore  -3.0629e-02  2.0321e-02 -1.5072e+00  0.131756
6:NrFjali             -2.8535e+00  1.7449e-01 -1.6353e+01 < 2.2e-16 ***
6:NrShenjapikesimi     1.1511e+00  1.1386e-01  1.0110e+01 < 2.2e-16 ***
6:NrQe                -2.1107e+00  2.2433e-01 -9.4087e+00 < 2.2e-16 ***
6:NrSE                 7.0386e+00  3.9697e-01  1.7731e+01 < 2.2e-16 ***
```

```
7:(Intercept)          1.2977e+01  4.5282e-03  2.8657e+03 < 2.2e-16 ***
7:NrFjale              1.1162e-01  3.5988e-02  3.1017e+00  0.001924 **
7:NrBashketingellore  -6.0412e-02  1.0634e-02 -5.6811e+00 1.339e-08 ***
7:NrFjali             -2.2229e+00  1.0827e-01 -2.0530e+01 < 2.2e-16 ***
7:NrShenjapikesimi     1.7007e+00  6.7531e-02  2.5184e+01 < 2.2e-16 ***
7:NrQe                -2.2090e+00  1.2502e-01 -1.7669e+01 < 2.2e-16 ***
7:NrSE                 1.8767e+00  1.6440e-01  1.1415e+01 < 2.2e-16 ***
8:(Intercept)          1.3151e+01  3.9426e-03  3.3355e+03 < 2.2e-16 ***
8:NrFjale              9.5812e-02  3.5077e-02  2.7315e+00  0.006305 **
8:NrBashketingellore  -5.7034e-02  1.0352e-02 -5.5095e+00 3.599e-08 ***
8:NrFjali             -2.1100e+00  1.0605e-01 -1.9895e+01 < 2.2e-16 ***
8:NrShenjapikesimi     1.7103e+00  6.9601e-02  2.4572e+01 < 2.2e-16 ***
8:NrQe                -2.2799e+00  1.2157e-01 -1.8754e+01 < 2.2e-16 ***
8:NrSE                 1.9556e+00  1.6744e-01  1.1679e+01 < 2.2e-16 ***
9:(Intercept)          1.5087e+01  2.2919e-02  6.5829e+02 < 2.2e-16 ***
9:NrFjale              1.5870e-01  3.9095e-02  4.0593e+00 4.921e-05 ***
9:NrBashketingellore  -7.9196e-02  1.1986e-02 -6.6075e+00 3.909e-11 ***
9:NrFjali             -2.3790e+00  1.2277e-01 -1.9378e+01 < 2.2e-16 ***
9:NrShenjapikesimi     1.7658e+00  8.2786e-02  2.1329e+01 < 2.2e-16 ***
9:NrQe                -1.9937e+00  1.4046e-01 -1.4194e+01 < 2.2e-16 ***
9:NrSE                 1.8239e+00  1.5235e-01  1.1971e+01 < 2.2e-16 ***
10:(Intercept)         5.5376e+00  1.2047e-02  4.5968e+02 < 2.2e-16 ***
10:NrFjale             1.5473e-01  3.8019e-02  4.0697e+00 4.707e-05 ***
10:NrBashketingellore -7.4721e-02  1.1474e-02 -6.5122e+00 7.405e-11 ***
10:NrFjali            -2.3605e+00  1.1352e-01 -2.0793e+01 < 2.2e-16 ***
10:NrShenjapikesimi    1.7839e+00  7.4842e-02  2.3836e+01 < 2.2e-16 ***
10:NrQe               -2.1726e+00  1.3960e-01 -1.5563e+01 < 2.2e-16 ***
10:NrSE                1.8322e+00  1.8533e-01  9.8864e+00 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> Pred4<-predict(MLARTSH4,DATAARTSH)
> PredMult4<-matrix(Pred4,nrow=10,byrow=TRUE)
> PredMult4
      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
 [1,] "1"  "1"  "1"  "1"  "1"  "1"  "1"  "1"  "1"  "1"
 [2,] "2"  "2"  "2"  "2"  "2"  "2"  "1"  "2"  "2"  "2"
 [3,] "3"  "3"  "3"  "3"  "3"  "3"  "3"  "3"  "3"  "3"
 [4,] "4"  "4"  "4"  "4"  "4"  "4"  "4"  "4"  "4"  "4"
 [5,] "5"  "9"  "5"  "1"  "5"  "5"  "5"  "5"  "5"  "5"
 [6,] "6"  "6"  "6"  "6"  "6"  "6"  "6"  "6"  "6"  "6"
 [7,] "9"  "2"  "7"  "7"  "7"  "7"  "7"  "7"  "8"  "7"
 [8,] "8"  "8"  "8"  "8"  "10" "7"  "7"  "7"  "8"  "8"
 [9,] "10" "9"  "9"  "9"  "9"  "9"  "8"  "10" "9"  "9"
[10,] "10" "5"  "9"  "7"  "10" "8"  "9"  "10" "10" "10"
```

Rezultatet e matricës PredMult4 janë përmbledhur në tabelën 3.12. Saktësia e klasifikimit të teksteve në modelin MLARTSH4 është në 82 %. Vërehet që parametrat janë thuajse të gjithë statistikisht të rendësishëm me α=0.001.

AUTORI	Klasifikimi i sakte në %
1	100
2	90
3	100
4	100
5	80
6	100
7	70
8	60
9	70
10	50
Mesatarja	82

TABELA 3. 12: REZULTATET E KLASIFIKIMIT NE MODELIN MLARTSH4

Po të përjashtohen variabla të tjerë në model vërehet një renie të ndjeshme të probabilitetit të klasifikimit të sakte të teksteve. Keshtu qe modeli MLARTSH4 përcaktohet si modeli më i mirë.

Me poshtë po shtojme variablat B_Z dhe G_F në model.

```
> MLARTSH5<-multinom(AUTORI~NrFjale+NrGerma+NrBashketingellore+NrZanore+B_Z+G_F+NrFjali+N
rShenjapikesimi+NrQe+NrSE,data=DATAARTSH)
# weights:  120 (99 variable)
initial  value 230.258509
iter  10 value 211.946888
iter  20 value 153.906904
iter  30 value 137.584697
iter  40 value 123.368772
iter  50 value 114.152551
iter  60 value 92.410166
iter  70 value 73.043660
iter  80 value 54.839251
iter  90 value 48.887260
iter 100 value 43.705222
final  value 43.705222
stopped after 100 iterations
> summary(MLARTSH5)
Call:
multinom(formula = AUTORI ~ NrFjale + NrGerma + NrBashketingellore +
```

```
    NrZanore + B_Z + G_F + NrFjali + NrShenjapikesimi + NrQe +
    NrSE, data = DATAARTSH)

Coefficients:
   (Intercept)     NrFjale     NrGerma NrBashketingellore     NrZanore        B_Z
 G_F   NrFjali
2     7.017509 -0.09123478 -0.01651867         0.13269076 -0.149209454 -30.467716 -23.710
0436 -3.317204
3     8.447435  0.86586857 -0.13105783         0.01780035 -0.148858206  15.192453  10.947
5386 -8.453652
4    -2.345219  0.91159797 -0.10383774        -0.14826728  0.044429558  -2.650907 -16.595
3628 -3.118031
5   -11.802516  0.66692183 -0.09664389        -0.06584506 -0.030798805   4.230457   0.145
1141 -4.402054
6    -1.547645  0.36893342 -0.03454114        -0.23846829  0.203927132  15.295764 -20.187
3823 -4.154917
7     4.170196  0.51854824 -0.07894963        -0.05229051 -0.026659167  -9.773944   2.472
9538 -4.179504
8    -6.332374  0.51819247 -0.08178621        -0.04122275 -0.040563434 -26.017733   9.851
3613 -4.030015
9     7.041433  0.61761499 -0.08985518        -0.07383996 -0.016015253  10.180718  -5.057
5373 -4.546462
10   -2.109015  0.58373612 -0.08220873        -0.07992589 -0.002282851  19.877029  -9.186
5059 -4.498894
   NrShenjapikesimi      NrQe        NrSE
2          3.188569 -6.837049 11.47267127
3          2.625801 -4.263305  6.76824847
4          2.593093 -7.699041  0.08350083
5          2.930071 -5.187059  6.02711155
6          2.697525 -4.588743 12.26108219
7          2.958981 -5.326022  6.00426260
8          2.974665 -5.433976  6.11320284
9          3.034254 -5.023656  5.91050078
10         3.079704 -5.269999  5.69359989

Std. Errors:
    (Intercept)    NrFjale     NrGerma NrBashketingellore   NrZanore          B_Z
  G_F    NrFjali
2  0.0021234327 0.10310552 0.011920441       0.0157199790 0.01638749 0.0028888125 1.04638
3e-02 0.31015621
3  0.0007129555 0.11525163 0.019006601       0.0247345001 0.03289734 0.0010247285 3.88151
4e-03 0.03789645
4  0.0001878221 0.23980122 0.032623315       0.0007919769 0.03284358 0.0002913061 7.67477
1e-05 0.03965623
5  0.0030760811 0.07252223 0.009906138       0.0235347162 0.02423444 0.0045221384 1.43851
9e-02 0.08470248
6  0.0051302321 0.05853768 0.009423113       0.0454824725 0.05023469 0.0075795072 2.43400
5e-02 0.07084011
7  0.0011828986 0.07598492 0.010038330       0.0241012114 0.02435772 0.0017047408 5.69572
8e-03 0.09030171
8  0.0028053119 0.07833365 0.010252588       0.0244725915 0.02468121 0.0040848831 1.37231
0e-02 0.08994943
9  0.0042796691 0.07086110 0.009656937       0.0240613497 0.02469961 0.0064686419 2.06937
1e-02 0.09405261
```

```
10 0.0056191114 0.07174329 0.009717554       0.0246118822 0.02529370 0.0083214928 2.67399
0e-02 0.07720382
   NrShenjapikesimi        NrQe        NrSE
2        0.22222287 0.081378429 0.053599363
3        0.16535491 0.076861148 0.082297386
4        0.08071695 0.009388023 0.005782606
5        0.07213308 0.097542862 0.144647924
6        0.09939062 0.140067116 0.057119089
7        0.07748462 0.104550716 0.141179863
8        0.07666168 0.112187824 0.162819456
9        0.06549790 0.095875580 0.120155160
10       0.06484592 0.112518974 0.164642611

Residual Deviance: 87.41044
AIC: 267.4104
> coeftest(MLARTSH5)

z test of coefficients:

                         Estimate  Std. Error     z value  Pr(>|z|)
2:(Intercept)          7.0175e+00  2.1234e-03  3.3048e+03 < 2.2e-16 ***
2:NrFjale             -9.1235e-02  1.0311e-01 -8.8490e-01 0.3762278
2:NrGerma             -1.6519e-02  1.1920e-02 -1.3857e+00 0.1658254
2:NrBashketingellore   1.3269e-01  1.5720e-02  8.4409e+00 < 2.2e-16 ***
2:NrZanore            -1.4921e-01  1.6387e-02 -9.1051e+00 < 2.2e-16 ***
2:B_Z                 -3.0468e+01  2.8888e-03 -1.0547e+04 < 2.2e-16 ***
2:G_F                 -2.3710e+01  1.0464e-02 -2.2659e+03 < 2.2e-16 ***
2:NrFjali             -3.3172e+00  3.1016e-01 -1.0695e+01 < 2.2e-16 ***
2:NrShenjapikesimi     3.1886e+00  2.2222e-01  1.4348e+01 < 2.2e-16 ***
2:NrQe                -6.8370e+00  8.1378e-02 -8.4016e+01 < 2.2e-16 ***
2:NrSE                 1.1473e+01  5.3599e-02  2.1404e+02 < 2.2e-16 ***
3:(Intercept)          8.4474e+00  7.1296e-04  1.1848e+04 < 2.2e-16 ***
3:NrFjale              8.6587e-01  1.1525e-01  7.5129e+00 5.785e-14 ***
3:NrGerma             -1.3106e-01  1.9007e-02 -6.8954e+00 5.372e-12 ***
3:NrBashketingellore   1.7800e-02  2.4734e-02  7.1970e-01 0.4717363
3:NrZanore            -1.4886e-01  3.2897e-02 -4.5249e+00 6.042e-06 ***
3:B_Z                  1.5192e+01  1.0247e-03  1.4826e+04 < 2.2e-16 ***
3:G_F                  1.0948e+01  3.8815e-03  2.8204e+03 < 2.2e-16 ***
3:NrFjali             -8.4537e+00  3.7896e-02 -2.2307e+02 < 2.2e-16 ***
3:NrShenjapikesimi     2.6258e+00  1.6535e-01  1.5880e+01 < 2.2e-16 ***
3:NrQe                -4.2633e+00  7.6861e-02 -5.5468e+01 < 2.2e-16 ***
3:NrSE                 6.7682e+00  8.2297e-02  8.2241e+01 < 2.2e-16 ***
4:(Intercept)         -2.3452e+00  1.8782e-04 -1.2486e+04 < 2.2e-16 ***
4:NrFjale              9.1160e-01  2.3980e-01  3.8015e+00 0.0001438 ***
4:NrGerma             -1.0384e-01  3.2623e-02 -3.1829e+00 0.0014579 **
4:NrBashketingellore  -1.4827e-01  7.9198e-04 -1.8721e+02 < 2.2e-16 ***
4:NrZanore             4.4430e-02  3.2844e-02  1.3528e+00 0.1761316
4:B_Z                 -2.6509e+00  2.9131e-04 -9.1001e+03 < 2.2e-16 ***
4:G_F                 -1.6595e+01  7.6748e-05 -2.1623e+05 < 2.2e-16 ***
4:NrFjali             -3.1180e+00  3.9656e-02 -7.8626e+01 < 2.2e-16 ***
4:NrShenjapikesimi     2.5931e+00  8.0717e-02  3.2126e+01 < 2.2e-16 ***
4:NrQe                -7.6990e+00  9.3880e-03 -8.2009e+02 < 2.2e-16 ***
4:NrSE                 8.3501e-02  5.7826e-03  1.4440e+01 < 2.2e-16 ***
5:(Intercept)         -1.1803e+01  3.0761e-03 -3.8369e+03 < 2.2e-16 ***
5:NrFjale              6.6692e-01  7.2522e-02  9.1961e+00 < 2.2e-16 ***
```

```
5:NrGerma             -9.6644e-02  9.9061e-03 -9.7560e+00 < 2.2e-16 ***
5:NrBashketingellore  -6.5845e-02  2.3535e-02 -2.7978e+00 0.0051454 **
5:NrZanore            -3.0799e-02  2.4234e-02 -1.2709e+00 0.2037753
5:B_Z                  4.2305e+00  4.5221e-03  9.3550e+02 < 2.2e-16 ***
5:G_F                  1.4511e-01  1.4385e-02  1.0088e+01 < 2.2e-16 ***
5:NrFjali             -4.4021e+00  8.4702e-02 -5.1971e+01 < 2.2e-16 ***
5:NrShenjapikesimi     2.9301e+00  7.2133e-02  4.0620e+01 < 2.2e-16 ***
5:NrQe                -5.1871e+00  9.7543e-02 -5.3177e+01 < 2.2e-16 ***
5:NrSE                 6.0271e+00  1.4465e-01  4.1667e+01 < 2.2e-16 ***
6:(Intercept)         -1.5476e+00  5.1302e-03 -3.0167e+02 < 2.2e-16 ***
6:NrFjale              3.6893e-01  5.8538e-02  6.3025e+00 2.929e-10 ***
6:NrGerma             -3.4541e-02  9.4231e-03 -3.6656e+00 0.0002468 ***
6:NrBashketingellore  -2.3847e-01  4.5482e-02 -5.2431e+00 1.579e-07 ***
6:NrZanore             2.0393e-01  5.0235e-02  4.0595e+00 4.918e-05 ***
6:B_Z                  1.5296e+01  7.5795e-03  2.0180e+03 < 2.2e-16 ***
6:G_F                 -2.0187e+01  2.4340e-02 -8.2939e+02 < 2.2e-16 ***
6:NrFjali             -4.1549e+00  7.0840e-02 -5.8652e+01 < 2.2e-16 ***
6:NrShenjapikesimi     2.6975e+00  9.9391e-02  2.7141e+01 < 2.2e-16 ***
6:NrQe                -4.5887e+00  1.4007e-01 -3.2761e+01 < 2.2e-16 ***
6:NrSE                 1.2261e+01  5.7119e-02  2.1466e+02 < 2.2e-16 ***
7:(Intercept)          4.1702e+00  1.1829e-03  3.5254e+03 < 2.2e-16 ***
7:NrFjale              5.1855e-01  7.5985e-02  6.8244e+00 8.832e-12 ***
7:NrGerma             -7.8950e-02  1.0038e-02 -7.8648e+00 3.696e-15 ***
7:NrBashketingellore  -5.2291e-02  2.4101e-02 -2.1696e+00 0.0300355 *
7:NrZanore            -2.6659e-02  2.4358e-02 -1.0945e+00 0.2737421
7:B_Z                 -9.7739e+00  1.7047e-03 -5.7334e+03 < 2.2e-16 ***
7:G_F                  2.4730e+00  5.6957e-03  4.3418e+02 < 2.2e-16 ***
7:NrFjali             -4.1795e+00  9.0302e-02 -4.6284e+01 < 2.2e-16 ***
7:NrShenjapikesimi     2.9590e+00  7.7485e-02  3.8188e+01 < 2.2e-16 ***
7:NrQe                -5.3260e+00  1.0455e-01 -5.0942e+01 < 2.2e-16 ***
7:NrSE                 6.0043e+00  1.4118e-01  4.2529e+01 < 2.2e-16 ***
8:(Intercept)         -6.3324e+00  2.8053e-03 -2.2573e+03 < 2.2e-16 ***
8:NrFjale              5.1819e-01  7.8334e-02  6.6152e+00 3.711e-11 ***
8:NrGerma             -8.1786e-02  1.0253e-02 -7.9771e+00 1.498e-15 ***
8:NrBashketingellore  -4.1223e-02  2.4473e-02 -1.6844e+00 0.0920955 .
8:NrZanore            -4.0563e-02  2.4681e-02 -1.6435e+00 0.1002807
8:B_Z                 -2.6018e+01  4.0849e-03 -6.3693e+03 < 2.2e-16 ***
8:G_F                  9.8514e+00  1.3723e-02  7.1787e+02 < 2.2e-16 ***
8:NrFjali             -4.0300e+00  8.9949e-02 -4.4803e+01 < 2.2e-16 ***
8:NrShenjapikesimi     2.9747e+00  7.6662e-02  3.8803e+01 < 2.2e-16 ***
8:NrQe                -5.4340e+00  1.1219e-01 -4.8436e+01 < 2.2e-16 ***
8:NrSE                 6.1132e+00  1.6282e-01  3.7546e+01 < 2.2e-16 ***
9:(Intercept)          7.0414e+00  4.2797e-03  1.6453e+03 < 2.2e-16 ***
9:NrFjale              6.1761e-01  7.0861e-02  8.7159e+00 < 2.2e-16 ***
9:NrGerma             -8.9855e-02  9.6569e-03 -9.3047e+00 < 2.2e-16 ***
9:NrBashketingellore  -7.3840e-02  2.4061e-02 -3.0688e+00 0.0021491 **
9:NrZanore            -1.6015e-02  2.4700e-02 -6.4840e-01 0.5167255
9:B_Z                  1.0181e+01  6.4686e-03  1.5739e+03 < 2.2e-16 ***
9:G_F                 -5.0575e+00  2.0694e-02 -2.4440e+02 < 2.2e-16 ***
9:NrFjali             -4.5465e+00  9.4053e-02 -4.8340e+01 < 2.2e-16 ***
9:NrShenjapikesimi     3.0343e+00  6.5498e-02  4.6326e+01 < 2.2e-16 ***
9:NrQe                -5.0237e+00  9.5876e-02 -5.2398e+01 < 2.2e-16 ***
9:NrSE                 5.9105e+00  1.2016e-01  4.9191e+01 < 2.2e-16 ***
10:(Intercept)        -2.1090e+00  5.6191e-03 -3.7533e+02 < 2.2e-16 ***
10:NrFjale             5.8374e-01  7.1743e-02  8.1365e+00 4.070e-16 ***
```

```
10:NrGerma            -8.2209e-02  9.7176e-03 -8.4598e+00 < 2.2e-16 ***
10:NrBashketingellore -7.9926e-02  2.4612e-02 -3.2475e+00 0.0011644 **
10:NrZanore           -2.2829e-03  2.5294e-02 -9.0300e-02 0.9280856
10:B_Z                 1.9877e+01  8.3215e-03  2.3886e+03 < 2.2e-16 ***
10:G_F                -9.1865e+00  2.6740e-02 -3.4355e+02 < 2.2e-16 ***
10:NrFjali            -4.4989e+00  7.7204e-02 -5.8273e+01 < 2.2e-16 ***
10:NrShenjapikesimi    3.0797e+00  6.4846e-02  4.7493e+01 < 2.2e-16 ***
10:NrQe               -5.2700e+00  1.1252e-01 -4.6837e+01 < 2.2e-16 ***
10:NrSE                5.6936e+00  1.6464e-01  3.4582e+01 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> Pred5<-predict(MLARTSH5,DATAARTSH)
> PredMult5<-matrix(Pred5,nrow=10,byrow=TRUE)
> PredMult5
      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
 [1,] "1"  "1"  "1"  "1"  "1"  "1"  "1"  "1"  "1"  "1"
 [2,] "2"  "2"  "2"  "2"  "2"  "2"  "2"  "2"  "2"  "2"
 [3,] "3"  "3"  "3"  "3"  "3"  "3"  "3"  "3"  "3"  "3"
 [4,] "4"  "4"  "4"  "4"  "4"  "4"  "4"  "4"  "4"  "4"
 [5,] "5"  "9"  "5"  "5"  "5"  "5"  "5"  "5"  "9"  "5"
 [6,] "6"  "6"  "6"  "6"  "6"  "6"  "6"  "6"  "6"  "6"
 [7,] "9"  "8"  "7"  "7"  "7"  "7"  "7"  "6"  "8"  "7"
 [8,] "8"  "8"  "8"  "8"  "10" "7"  "7"  "7"  "8"  "8"
 [9,] "10" "10" "7"  "9"  "9"  "9"  "9"  "10" "10" "9"
[10,] "10" "5"  "9"  "7"  "10" "10" "9"  "10" "10" "10"
```

AUTORI	Klasifikimi i sakte në %
1	100
2	100
3	100
4	100
5	80
6	100
7	60
8	60
9	50
10	60
Mesatarja	81

TABELA 3. 13: REZULTATET E KLASIFIKIMIT NE MODELIN MLARTSH5

Rezultatet e matricës PredMult5 janë përmbledhur në tabelën 3.12. Saktësia e klasifikimit të teksteve në modelin MLARTSH5 është në 81 %, por variabli NrZanore nuk është statistikisht i rëndësishëm.

Në modelin më poshtë përjashtohet variabli NrZanore.

```
> MLARTSH6<-multinom(AUTORI~NrFjale+NrGerma+NrBashketingellore+B_Z+G_F+NrFjali+NrShenjapi
kesimi+NrQe+NrSE,data=DATAARTSH)
# weights:  110 (90 variable)
initial  value 230.258509
iter  10 value 200.740032
iter  20 value 131.565634
iter  30 value 116.573136
iter  40 value 105.242398
iter  50 value 99.693544
iter  60 value 84.047855
iter  70 value 68.575751
iter  80 value 52.781613
iter  90 value 48.510289
iter 100 value 43.988541
final  value 43.988541
stopped after 100 iterations
> summary(MLARTSH6)
Call:
multinom(formula = AUTORI ~ NrFjale + NrGerma + NrBashketingellore +
    B_Z + G_F + NrFjali + NrShenjapikesimi + NrQe + NrSE, data = DATAARTSH)

Coefficients:
   (Intercept)     NrFjale     NrGerma NrBashketingellore        B_Z         G_F   NrFjal
i NrShenjapikesimi
2    7.0182735 -0.03675842 -0.14957458        0.232662614 -27.027504 -16.9381789 -2.90384
3         2.839282
3    3.8491643  0.90519552 -0.25983020        0.113168382   5.897223   9.7846890 -7.56231
3         2.631370
4   -2.4594564  0.92371277 -0.03288124       -0.241209386  -3.637058  -9.0700590 -2.73840
8         2.180117
5  -11.2197474  0.63622991 -0.11709581       -0.035628503   5.960651  -0.2859873 -3.94230
1         2.450490
6   -0.9166951  0.45906970  0.20346133       -0.530563898  10.161882 -14.3878285 -4.08555
0         2.721916
7    4.3978578  0.50615117 -0.09441198       -0.031899108  -3.469355  -0.5070670 -3.83025
2         2.500855
8   -4.2442364  0.50464103 -0.11147506       -0.005830693 -21.187019   6.8693896 -3.69176
2         2.514194
9    5.6323334  0.59592506 -0.10036961       -0.054089951  10.239432  -4.2124562 -4.13363
9         2.576811
10   0.3435848  0.54755694 -0.07500668       -0.074800495  23.980342 -10.5958734 -4.06076
8         2.602368
        NrQe       NrSE
2  -5.719047  9.8658733
3  -4.004815  6.5176829
4  -7.755091 -0.3370724
5  -4.324592  5.5513685
6  -4.118659  9.5597885
7  -4.380112  5.5406460
8  -4.466085  5.6719477
9  -4.175608  5.4122316
10 -4.409366  5.2192356
```

```
Std. Errors:
    (Intercept)    NrFjale    NrGerma NrBashketingellore          B_Z         G_F    NrFja
li NrShenjapikesimi
2  0.0027555645 0.1336267 0.02807809         0.01566337 0.0037849134 0.013848909 0.202794
67      0.173455061
3  0.0013669942 0.1931111 0.09641442         0.15129219 0.0020733199 0.007450982 0.022399
82      0.068027716
4  0.0005866007 0.1893200 0.02210415         0.02705557 0.0008336041 0.002005928 0.003619
08      0.008428569
5  0.0096993051 0.1037141 0.02011817         0.03096810 0.0142370695 0.045819901 0.084281
55      0.099397193
6  0.0218692244 0.1080271 0.03417508         0.05938899 0.0319596929 0.106355313 0.119241
68      0.123132176
7  0.0016052067 0.1017098 0.01933999         0.03168522 0.0023935043 0.008122344 0.082515
31      0.095357997
8  0.0020658826 0.1038389 0.02003804         0.03289390 0.0030458261 0.010068927 0.086386
26      0.096502979
9  0.0106630038 0.1028604 0.02027417         0.03212677 0.0159520371 0.051150392 0.091106
60      0.095636740
10 0.0058921024 0.1004339 0.02011110         0.03349991 0.0085063538 0.029676169 0.085792
11      0.094164666
         NrQe         NrSE
2  0.10804862 0.0544687302
3  0.08553792 0.0455795107
4  0.01487883 0.0009732688
5  0.10209727 0.1418637911
6  0.13799726 0.2678035734
7  0.10673268 0.1508396249
8  0.11428748 0.1703230735
9  0.09884955 0.1257499892
10 0.10386836 0.1718219059

Residual Deviance: 87.97708
AIC: 267.9771
> coeftest(MLARTSH6)

z test of coefficients:

                         Estimate  Std. Error    z value  Pr(>|z|)
2:(Intercept)          7.0183e+00  2.7556e-03  2546.9458 < 2.2e-16 ***
2:NrFjale             -3.6758e-02  1.3363e-01    -0.2751 0.7832525
2:NrGerma             -1.4957e-01  2.8078e-02    -5.3271 9.980e-08 ***
2:NrBashketingellore   2.3266e-01  1.5663e-02    14.8539 < 2.2e-16 ***
2:B_Z                 -2.7028e+01  3.7849e-03 -7140.8514 < 2.2e-16 ***
2:G_F                 -1.6938e+01  1.3849e-02 -1223.0696 < 2.2e-16 ***
2:NrFjali             -2.9038e+00  2.0279e-01   -14.3191 < 2.2e-16 ***
2:NrShenjapikesimi     2.8393e+00  1.7346e-01    16.3690 < 2.2e-16 ***
2:NrQe                -5.7190e+00  1.0805e-01   -52.9303 < 2.2e-16 ***
2:NrSE                 9.8659e+00  5.4469e-02   181.1291 < 2.2e-16 ***
3:(Intercept)          3.8492e+00  1.3670e-03  2815.7869 < 2.2e-16 ***
3:NrFjale              9.0520e-01  1.9311e-01     4.6874 2.767e-06 ***
3:NrGerma             -2.5983e-01  9.6414e-02    -2.6949 0.0070403 **
3:NrBashketingellore   1.1317e-01  1.5129e-01     0.7480 0.4544529
3:B_Z                  5.8972e+00  2.0733e-03  2844.3382 < 2.2e-16 ***
```

```
3:G_F                 9.7847e+00  7.4510e-03  1313.2080 < 2.2e-16 ***
3:NrFjali            -7.5623e+00  2.2400e-02  -337.6061 < 2.2e-16 ***
3:NrShenjapikesimi    2.6314e+00  6.8028e-02    38.6809 < 2.2e-16 ***
3:NrQe               -4.0048e+00  8.5538e-02   -46.8192 < 2.2e-16 ***
3:NrSE                6.5177e+00  4.5580e-02   142.9959 < 2.2e-16 ***
4:(Intercept)        -2.4595e+00  5.8660e-04 -4192.7264 < 2.2e-16 ***
4:NrFjale             9.2371e-01  1.8932e-01     4.8791 1.066e-06 ***
4:NrGerma            -3.2881e-02  2.2104e-02    -1.4876 0.1368670
4:NrBashketingellore -2.4121e-01  2.7056e-02    -8.9153 < 2.2e-16 ***
4:B_Z                -3.6371e+00  8.3360e-04 -4363.0519 < 2.2e-16 ***
4:G_F                -9.0701e+00  2.0059e-03 -4521.6273 < 2.2e-16 ***
4:NrFjali            -2.7384e+00  3.6191e-03  -756.6586 < 2.2e-16 ***
4:NrShenjapikesimi    2.1801e+00  8.4286e-03   258.6580 < 2.2e-16 ***
4:NrQe               -7.7551e+00  1.4879e-02  -521.2166 < 2.2e-16 ***
4:NrSE               -3.3707e-01  9.7327e-04  -346.3302 < 2.2e-16 ***
5:(Intercept)        -1.1220e+01  9.6993e-03 -1156.7579 < 2.2e-16 ***
5:NrFjale             6.3623e-01  1.0371e-01     6.1345 8.545e-10 ***
5:NrGerma            -1.1710e-01  2.0118e-02    -5.8204 5.871e-09 ***
5:NrBashketingellore -3.5629e-02  3.0968e-02    -1.1505 0.2499420
5:B_Z                 5.9607e+00  1.4237e-02   418.6712 < 2.2e-16 ***
5:G_F                -2.8599e-01  4.5820e-02    -6.2416 4.333e-10 ***
5:NrFjali            -3.9423e+00  8.4282e-02   -46.7754 < 2.2e-16 ***
5:NrShenjapikesimi    2.4505e+00  9.9397e-02    24.6535 < 2.2e-16 ***
5:NrQe               -4.3246e+00  1.0210e-01   -42.3576 < 2.2e-16 ***
5:NrSE                5.5514e+00  1.4186e-01    39.1317 < 2.2e-16 ***
6:(Intercept)        -9.1670e-01  2.1869e-02   -41.9171 < 2.2e-16 ***
6:NrFjale             4.5907e-01  1.0803e-01     4.2496 2.142e-05 ***
6:NrGerma             2.0346e-01  3.4175e-02     5.9535 2.625e-09 ***
6:NrBashketingellore -5.3056e-01  5.9389e-02    -8.9337 < 2.2e-16 ***
6:B_Z                 1.0162e+01  3.1960e-02   317.9593 < 2.2e-16 ***
6:G_F                -1.4388e+01  1.0636e-01  -135.2808 < 2.2e-16 ***
6:NrFjali            -4.0855e+00  1.1924e-01   -34.2628 < 2.2e-16 ***
6:NrShenjapikesimi    2.7219e+00  1.2313e-01    22.1056 < 2.2e-16 ***
6:NrQe               -4.1187e+00  1.3800e-01   -29.8459 < 2.2e-16 ***
6:NrSE                9.5598e+00  2.6780e-01    35.6970 < 2.2e-16 ***
7:(Intercept)         4.3979e+00  1.6052e-03  2739.7455 < 2.2e-16 ***
7:NrFjale             5.0615e-01  1.0171e-01     4.9764 6.477e-07 ***
7:NrGerma            -9.4412e-02  1.9340e-02    -4.8817 1.052e-06 ***
7:NrBashketingellore -3.1899e-02  3.1685e-02    -1.0068 0.3140547
7:B_Z                -3.4694e+00  2.3935e-03 -1449.4875 < 2.2e-16 ***
7:G_F                -5.0707e-01  8.1223e-03   -62.4287 < 2.2e-16 ***
7:NrFjali            -3.8303e+00  8.2515e-02   -46.4187 < 2.2e-16 ***
7:NrShenjapikesimi    2.5009e+00  9.5358e-02    26.2260 < 2.2e-16 ***
7:NrQe               -4.3801e+00  1.0673e-01   -41.0381 < 2.2e-16 ***
7:NrSE                5.5406e+00  1.5084e-01    36.7320 < 2.2e-16 ***
8:(Intercept)        -4.2442e+00  2.0659e-03 -2054.4422 < 2.2e-16 ***
8:NrFjale             5.0464e-01  1.0384e-01     4.8598 1.175e-06 ***
8:NrGerma            -1.1148e-01  2.0038e-02    -5.5632 2.649e-08 ***
8:NrBashketingellore -5.8307e-03  3.2894e-02    -0.1773 0.8593061
8:B_Z                -2.1187e+01  3.0458e-03 -6956.0828 < 2.2e-16 ***
8:G_F                 6.8694e+00  1.0069e-02   682.2365 < 2.2e-16 ***
8:NrFjali            -3.6918e+00  8.6386e-02   -42.7355 < 2.2e-16 ***
8:NrShenjapikesimi    2.5142e+00  9.6503e-02    26.0530 < 2.2e-16 ***
8:NrQe               -4.4661e+00  1.1429e-01   -39.0776 < 2.2e-16 ***
8:NrSE                5.6719e+00  1.7032e-01    33.3011 < 2.2e-16 ***
```

```
9:(Intercept)          5.6323e+00  1.0663e-02   528.2126 < 2.2e-16 ***
9:NrFjale              5.9593e-01  1.0286e-01     5.7935 6.892e-09 ***
9:NrGerma             -1.0037e-01  2.0274e-02    -4.9506 7.398e-07 ***
9:NrBashketingellore  -5.4090e-02  3.2127e-02    -1.6836 0.0922510 .
9:B_Z                  1.0239e+01  1.5952e-02   641.8887 < 2.2e-16 ***
9:G_F                 -4.2125e+00  5.1150e-02   -82.3543 < 2.2e-16 ***
9:NrFjali             -4.1336e+00  9.1107e-02   -45.3715 < 2.2e-16 ***
9:NrShenjapikesimi     2.5768e+00  9.5637e-02    26.9437 < 2.2e-16 ***
9:NrQe                -4.1756e+00  9.8850e-02   -42.2421 < 2.2e-16 ***
9:NrSE                 5.4122e+00  1.2575e-01    43.0396 < 2.2e-16 ***
10:(Intercept)         3.4358e-01  5.8921e-03    58.3128 < 2.2e-16 ***
10:NrFjale             5.4756e-01  1.0043e-01     5.4519 4.983e-08 ***
10:NrGerma            -7.5007e-02  2.0111e-02    -3.7296 0.0001918 ***
10:NrBashketingellore -7.4800e-02  3.3500e-02    -2.2329 0.0255584 *
10:B_Z                 2.3980e+01  8.5063e-03  2819.1094 < 2.2e-16 ***
10:G_F                -1.0596e+01  2.9676e-02  -357.0499 < 2.2e-16 ***
10:NrFjali            -4.0608e+00  8.5792e-02   -47.3326 < 2.2e-16 ***
10:NrShenjapikesimi    2.6024e+00  9.4165e-02    27.6364 < 2.2e-16 ***
10:NrQe               -4.4094e+00  1.0387e-01   -42.4515 < 2.2e-16 ***
10:NrSE                5.2192e+00  1.7182e-01    30.3758 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> Pred6<-predict(MLARTSH6,DATAARTSH)
> PredMult6<-matrix(Pred6,nrow=10,byrow=TRUE)
> PredMult6
      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
 [1,] "1"  "1"  "1"  "1"  "1"  "1"  "1"  "1"  "1"  "1"
 [2,] "2"  "2"  "2"  "2"  "2"  "2"  "2"  "2"  "2"  "2"
 [3,] "3"  "3"  "3"  "3"  "3"  "3"  "3"  "3"  "3"  "3"
 [4,] "4"  "4"  "4"  "4"  "4"  "4"  "4"  "4"  "4"  "4"
 [5,] "5"  "9"  "5"  "5"  "5"  "5"  "5"  "5"  "5"  "5"
 [6,] "6"  "6"  "6"  "6"  "6"  "6"  "6"  "6"  "6"  "6"
 [7,] "9"  "8"  "7"  "7"  "7"  "7"  "7"  "6"  "8"  "7"
 [8,] "8"  "8"  "8"  "8"  "7"  "7"  "7"  "7"  "8"  "8"
 [9,] "10" "9"  "7"  "9"  "9"  "9"  "8"  "10" "10" "9"
[10,] "10" "5"  "10" "7"  "10" "10" "9"  "10" "10" "10"
```

Rezultatet e matricës PredMult6 janë përmbledhur në tabelën 3.14. Nga rezultatet klasifikimi i saktë është bërë me probabilitet 0.83 dhe thuajse të gjithë variablat janë statistikisht te rëndesishem.

AUTORI	Klasifikimi i sakte në %
1	100
2	100
3	100
4	100
5	90
6	100
7	60
8	60
9	50
10	70
Mesatarja	83

TABELA 3. 14: REZULTATET E KLASIFIKIMIT NE MODELIN MLARTSH6

Në qoftë se reduktojmë variabla të tjerë kemi rënie të probabilitetit të klasifikimit të saktë të teksteve, sic tregohet edhe në modelin me poshtë.

```
> MLARTSH7<-multinom(AUTORI~NrBashketingellore+B_Z+G_F+NrFjali+NrShenjapikesimi+NrQe+NrSE
,data=DATAARTSH)
# weights:  90 (72 variable)
initial  value 230.258509
iter  10 value 194.740205
iter  20 value 144.001856
iter  30 value 133.334300
iter  40 value 124.104052
iter  50 value 104.479505
iter  60 value 75.176549
iter  70 value 69.348027
iter  80 value 63.167691
iter  90 value 58.817908
iter 100 value 55.637098
final  value 55.637098
stopped after 100 iterations
> summary(MLARTSH7)
Call:
multinom(formula = AUTORI ~ NrBashketingellore + B_Z + G_F +
    NrFjali + NrShenjapikesimi + NrQe + NrSE, data = DATAARTSH)

Coefficients:
   (Intercept) NrBashketingellore       B_Z         G_F    NrFjali NrShenjapikesimi
 NrQe      NrSE
2   -90.068651       -0.007613067  45.48224  -2.6864457 -0.9715921        0.9219969 -1.78
75931  2.426200
```

```
3     6.660356       -0.014560625  31.07351   8.9770620 -3.3698018        0.6316687 -0.67
83401  1.226974
4    21.282841       -0.003340472  18.42850 -15.5317261 -1.3052022        1.4306692 -3.92
05526 -4.003768
5   116.925529       -0.009274587  22.29757 -32.4157114 -1.2043534        0.8728164 -1.22
19148  1.710999
6   -14.012126       -0.009235782 -47.18118  10.5028151 -1.8166373        0.7815301 -0.85
83935  4.713059
7   -26.313169       -0.014993384  17.11613  -0.4215607 -1.1849820        0.9790509 -1.18
92288  1.600922
8   -31.800090       -0.016662105  19.66320  -0.2115559 -1.1217260        1.0093644 -1.26
86803  1.651614
9    83.064742       -0.014845654  16.97075 -22.4758328 -1.4195168        1.0236810 -1.05
99214  1.553945
10   62.895511       -0.011393610  17.61430 -20.4649861 -1.3819537        1.0322348 -1.27
59203  1.588863

Std. Errors:
    (Intercept) NrBashketingellore          B_Z          G_F    NrFjali NrShenjapikesimi
      NrQe
2  0.0896312852        0.002355599 0.1727867729 0.4483178023 0.09658093      0.070057732
0.139131784
3  0.0032796820        0.010521300 0.0047193262 0.0129077262 0.08965558      0.242725084
0.389856951
4  0.0002185138        0.006163701 0.0002889874 0.0007169336 0.02831128      0.006280275
0.001477951
5  0.0938332290        0.002888679 0.1602973028 0.4562343584 0.12628552      0.093916262
0.163857924
6  0.0980604125        0.004005913 0.1425853675 0.4996967389 0.17465951      0.122156457
0.235832128
7  0.0972856711        0.002370707 0.1487701225 0.4666704242 0.13156596      0.085663704
0.148155608
8  0.0945547421        0.002402975 0.1490859192 0.4655389496 0.12269986      0.082595202
0.147206214
9  0.1027265106        0.003036739 0.1517934813 0.5139904146 0.14629601      0.095178545
0.165640759
10 0.0886920849        0.002564392 0.1333674387 0.4260228898 0.14293974      0.088166693
0.144109285
           NrSE
2  0.2027706073
3  0.1911616716
4  0.0004377818
5  0.2042753623
6  0.4299149590
7  0.1881373919
8  0.1862006806
9  0.1911695469
10 0.1976326722

Residual Deviance: 111.2742
AIC: 255.2742
> coeftest(MLARTSH7)

z test of coefficients:
```

```
                        Estimate  Std. Error     z value  Pr(>|z|)
2:(Intercept)         -9.0069e+01  8.9631e-02  -1004.8796 < 2.2e-16 ***
2:NrBashketingellore  -7.6131e-03  2.3556e-03     -3.2319 0.0012297 **
2:B_Z                  4.5482e+01  1.7279e-01    263.2276 < 2.2e-16 ***
2:G_F                 -2.6864e+00  4.4832e-01     -5.9923 2.069e-09 ***
2:NrFjali             -9.7159e-01  9.6581e-02    -10.0599 < 2.2e-16 ***
2:NrShenjapikesimi     9.2200e-01  7.0058e-02     13.1605 < 2.2e-16 ***
2:NrQe                -1.7876e+00  1.3913e-01    -12.8482 < 2.2e-16 ***
2:NrSE                 2.4262e+00  2.0277e-01     11.9652 < 2.2e-16 ***
3:(Intercept)          6.6604e+00  3.2797e-03   2030.7931 < 2.2e-16 ***
3:NrBashketingellore  -1.4561e-02  1.0521e-02     -1.3839 0.1663833
3:B_Z                  3.1074e+01  4.7193e-03   6584.3114 < 2.2e-16 ***
3:G_F                  8.9771e+00  1.2908e-02    695.4797 < 2.2e-16 ***
3:NrFjali             -3.3698e+00  8.9656e-02    -37.5861 < 2.2e-16 ***
3:NrShenjapikesimi     6.3167e-01  2.4273e-01      2.6024 0.0092573 **
3:NrQe                -6.7834e-01  3.8986e-01     -1.7400 0.0818640 .
3:NrSE                 1.2270e+00  1.9116e-01      6.4185 1.376e-10 ***
4:(Intercept)          2.1283e+01  2.1851e-04  97398.1386 < 2.2e-16 ***
4:NrBashketingellore  -3.3405e-03  6.1637e-03     -0.5420 0.5878469
4:B_Z                  1.8429e+01  2.8899e-04  63769.2288 < 2.2e-16 ***
4:G_F                 -1.5532e+01  7.1693e-04 -21664.1082 < 2.2e-16 ***
4:NrFjali             -1.3052e+00  2.8311e-02    -46.1018 < 2.2e-16 ***
4:NrShenjapikesimi     1.4307e+00  6.2803e-03    227.8036 < 2.2e-16 ***
4:NrQe                -3.9206e+00  1.4780e-03  -2652.6953 < 2.2e-16 ***
4:NrSE                -4.0038e+00  4.3778e-04  -9145.5790 < 2.2e-16 ***
5:(Intercept)          1.1693e+02  9.3833e-02   1246.0994 < 2.2e-16 ***
5:NrBashketingellore  -9.2746e-03  2.8887e-03     -3.2107 0.0013243 **
5:B_Z                  2.2298e+01  1.6030e-01    139.1013 < 2.2e-16 ***
5:G_F                 -3.2416e+01  4.5623e-01    -71.0506 < 2.2e-16 ***
5:NrFjali             -1.2044e+00  1.2629e-01     -9.5368 < 2.2e-16 ***
5:NrShenjapikesimi     8.7282e-01  9.3916e-02      9.2936 < 2.2e-16 ***
5:NrQe                -1.2219e+00  1.6386e-01     -7.4572 8.841e-14 ***
5:NrSE                 1.7110e+00  2.0428e-01      8.3759 < 2.2e-16 ***
6:(Intercept)         -1.4012e+01  9.8060e-02   -142.8928 < 2.2e-16 ***
6:NrBashketingellore  -9.2358e-03  4.0059e-03     -2.3055 0.0211365 *
6:B_Z                 -4.7181e+01  1.4259e-01   -330.8977 < 2.2e-16 ***
6:G_F                  1.0503e+01  4.9970e-01     21.0184 < 2.2e-16 ***
6:NrFjali             -1.8166e+00  1.7466e-01    -10.4010 < 2.2e-16 ***
6:NrShenjapikesimi     7.8153e-01  1.2216e-01      6.3978 1.577e-10 ***
6:NrQe                -8.5839e-01  2.3583e-01     -3.6398 0.0002728 ***
6:NrSE                 4.7131e+00  4.2991e-01     10.9628 < 2.2e-16 ***
7:(Intercept)         -2.6313e+01  9.7286e-02   -270.4732 < 2.2e-16 ***
7:NrBashketingellore  -1.4993e-02  2.3707e-03     -6.3244 2.542e-10 ***
7:B_Z                  1.7116e+01  1.4877e-01    115.0508 < 2.2e-16 ***
7:G_F                 -4.2156e-01  4.6667e-01     -0.9033 0.3663470
7:NrFjali             -1.1850e+00  1.3157e-01     -9.0068 < 2.2e-16 ***
7:NrShenjapikesimi     9.7905e-01  8.5664e-02     11.4290 < 2.2e-16 ***
7:NrQe                -1.1892e+00  1.4816e-01     -8.0269 9.997e-16 ***
7:NrSE                 1.6009e+00  1.8814e-01      8.5093 < 2.2e-16 ***
8:(Intercept)         -3.1800e+01  9.4555e-02   -336.3141 < 2.2e-16 ***
8:NrBashketingellore  -1.6662e-02  2.4030e-03     -6.9339 4.093e-12 ***
8:B_Z                  1.9663e+01  1.4909e-01    131.8917 < 2.2e-16 ***
8:G_F                 -2.1156e-01  4.6554e-01     -0.4544 0.6495177
8:NrFjali             -1.1217e+00  1.2270e-01     -9.1420 < 2.2e-16 ***
8:NrShenjapikesimi     1.0094e+00  8.2595e-02     12.2206 < 2.2e-16 ***
```

```
8:NrQe                -1.2687e+00  1.4721e-01     -8.6184 < 2.2e-16 ***
8:NrSE                 1.6516e+00  1.8620e-01      8.8701 < 2.2e-16 ***
9:(Intercept)          8.3065e+01  1.0273e-01    808.6008 < 2.2e-16 ***
9:NrBashketingellore  -1.4846e-02  3.0367e-03     -4.8887 1.015e-06 ***
9:B_Z                  1.6971e+01  1.5179e-01    111.8015 < 2.2e-16 ***
9:G_F                 -2.2476e+01  5.1399e-01    -43.7281 < 2.2e-16 ***
9:NrFjali             -1.4195e+00  1.4630e-01     -9.7030 < 2.2e-16 ***
9:NrShenjapikesimi     1.0237e+00  9.5179e-02     10.7554 < 2.2e-16 ***
9:NrQe                -1.0599e+00  1.6564e-01     -6.3989 1.565e-10 ***
9:NrSE                 1.5539e+00  1.9117e-01      8.1286 4.342e-16 ***
10:(Intercept)         6.2896e+01  8.8692e-02    709.1446 < 2.2e-16 ***
10:NrBashketingellore -1.1394e-02  2.5644e-03     -4.4430 8.871e-06 ***
10:B_Z                 1.7614e+01  1.3337e-01    132.0734 < 2.2e-16 ***
10:G_F                -2.0465e+01  4.2602e-01    -48.0373 < 2.2e-16 ***
10:NrFjali            -1.3820e+00  1.4294e-01     -9.6681 < 2.2e-16 ***
10:NrShenjapikesimi    1.0322e+00  8.8167e-02     11.7078 < 2.2e-16 ***
10:NrQe               -1.2759e+00  1.4411e-01     -8.8538 < 2.2e-16 ***
10:NrSE                1.5889e+00  1.9763e-01      8.0395 9.023e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> Pred7<-predict(MLARTSH5,DATAARTSH)
> PredMult7<-matrix(Pred5,nrow=10,byrow=TRUE)
> PredMult7
      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
 [1,] "8"  "1"  "1"  "1"  "1"  "1"  "1"  "1"  "1"  "1"
 [2,] "2"  "2"  "2"  "2"  "2"  "2"  "1"  "2"  "2"  "2"
 [3,] "3"  "3"  "3"  "3"  "3"  "3"  "3"  "3"  "3"  "3"
 [4,] "4"  "4"  "4"  "4"  "4"  "4"  "4"  "4"  "4"  "4"
 [5,] "5"  "9"  "5"  "5"  "5"  "5"  "5"  "5"  "9"  "5"
 [6,] "6"  "6"  "6"  "6"  "6"  "6"  "6"  "6"  "6"  "6"
 [7,] "9"  "2"  "7"  "7"  "7"  "7"  "7"  "7"  "8"  "7"
 [8,] "8"  "7"  "8"  "8"  "10" "7"  "7"  "7"  "8"  "7"
 [9,] "10" "9"  "9"  "9"  "9"  "5"  "9"  "10" "9"  "9"
[10,] "10" "5"  "9"  "7"  "10" "10" "9"  "10" "2"  "10"
```

Vërehet që probabiliteti është ulur në 0.79 dhe variabli NrBashketingellore nuk është statistikisht i rëndësishëm. Në modelet e tjerë vërehet një ulje të ndjeshme minimumi deri në 69%. Te gjithe modelet e shqyrtuara janë paraqitur në Shtojca 2 (CD)

3.8 Përfundime

Në këtë kapitull u trajtua modeli i regresit logjistik si një metodë për klasifikimin e teksteve shqip me qëllim identifikimin e autorit. U shqyrtuan modele të ndryshme të regresit logjistik të shumëfishtë dhe të regresit logjistik multinomial duke konsideruar si variabla të pavarur të grumbulluar nga 100 tekste shqip të 10 autoreve të ndryshëm, numrin e fjalëve (NrFjale), numrin e germave (NrGerma), numrin e zanoreve (NrZanore), numrin e bashkëtingëlloreve (NrBashketingellore), numrin e shenjave të pikësimit (NrShenjapikesimi), numrin e fjalive, NrFjali, frekuencën e lidhëzës *që* (NrQE) dhe frekuencën e lidhëzës *se* (NrSE), këto dy variabla të fundit u grumbulluan meqenëse këto dy lidhëza janë ndër lidhëzat më të rëndësishme të cilat janë një klasë e rëndësishme gramatikore të fjalëve ne gjuhën shqipe. Meqenëse në kapitullin 2 u bënë vlerësime pikesore dhe vlerësime intervalore për dy koeficiente përkatësisht raporti i germave me fjalët G_F dhe raporti i bashkëtingëlloreve me zanoret B_Z, shtuam edhe ndikimin e këtyre dy variablave në modelet përkatëse.

Nga diskutimi i modeleve të ndërtuar u përcaktuan modelet statistikisht më të mirë, të cilët identifikojnë autorësinë e teksteve me probabilitetin më të lartë dhe përmbajnë parametrat statistikore më të mirë të këtyre variablave.

Nga shqyrtimi i modeleve të regresit logjistik të shumëfishtë u përcaktua modeli më i mire ai me variablat: NrFjali, NrShenjapikesimi, NrQË, B_Z dhe G_F i cili identifikon autorësinë e teksteve me një saktësi deri në 93.5%. Nga shqyrtimi i të gjithë modeleve të ndërtuar në Salillari et al.(2012), Salillari, Prifti (2016) dhe modeleve të trajtuara në këtë kapitull u arrit në përfundimin se modeli i regresit logjistik është një model shumë i mirë si metodë klasifikimi për tekstet shqip edhe duke konsideruar një numër jo shumë të madh të variablave të pavarur. Me shtimin e variablave të pavarur kemi rritje të probabilitetit të identifikimit të autorit të vërtetë nga 0.918 ne 0.935. E meta e këtij modeli qëndron në kohën e realizimit të tij sepse duhen ndërtuar aq modele sa autorë shqyrtohen në studim. Parametrat statistikisht më të rëndësishëm nuk janë të njëjtët në çdo model. Kjo gjë sjell vështirësi në përgjithësimin e parametrave më të mirë për të gjithë autorët, pavarësisht se probabiliteti i identifikimit të autorit të vërtetë të tekstit është shumë i madh.

Ndër modelet e regresit logjistik multinomial u përcaktuan si variabla më të rëndësishëm NrFjale, Nrbashketingellore, NrFjali, NrShenjapikesimi, NrQË dhe NrSE i cili identifikon autorësinë e teksteve me një saktësi deri në 83%. Edhe modeli kur shtojmë variablat B_Z dhe G_Z japin thuajse të njëjtin probabilitet të identifikimit të autorit të vërtetë, por jo të gjithë parametrat e modelit dalin statistikisht të rëndësishëm. Thuajse të gjithë modelet e shqyrtuara janë statistikisht të rëndësishëm me probabilitet të identifikimit të autorit të vërtetë të paktën 69%. Këto rezultate janë shumë më të mira se rezultatet e paraqitura në Salillari.D,Prifti (2016) në të cilin shqyrtuam më pak variabla të pavarur në model dhe probabiliteti i identifikimit të autorit të vërtetë rezultoi nga 0.59 deri ne 0.738. Krahasuar me modelet e regresit logjistik të shumëfishtë, në modelet e regresit logjistik multinomial përcaktohen variabla të ndryshëm statistikisht të rëndësishëm. Probabilitetet e identifikimit të autorit të vërtetë, pavarësisht se janë rezultate të mira, ato janë më të ulëta se rezultatet e përftuara në modelin e regresit

logjistik të shumëfishtë. Pavarësisht këtyre rezultateve modeli i regresit logjistik ka avantazh kundrejt modelit të regresit logjistik të shumëfishtë së pari elementin kohë për shkak se në regresin logjistik multinomial ndërtojmë vetëm një model ndërsa për modelin e regresit logjistik të shumëfishtë ndërtojmë aq modele sa autorë kemi dhe së dyti përgjithësimin e parametrave më të mirë të modelit.

Literatura

1. Alpaydın. E.(2010). Introduction to Machine Learning Second Edition, MIT p 157-160.
2. Baayen, R. H., H. van Halteren, and Tweedie, F. J., (1996). "Outside the cave of shadows. Using syntactic annotation to enhance authorship attribution", Literary and Linguistic Computing, , 11:121—131.
3. Baayen, R.H., Van Halteren, H., Neijt, A., & Tweedie, F. (2002). An expëriment in authorship attribution. In Proceedings of the 6th International Conference on the Statistical Analysis of Textual Data (JADT 2002).
4. Charu. C. Aggarwal, ChengXiang Zhai. (2012). A Survey of Text Clustering Algorithms, Springer Science+Business Media, LLC.
5. Çiço.B, Cepa.V, 2003. Toword a spellchecker for Albanian Language. BCI2003.
6. Dika. A, Maxhuni. A, Rexhepi.A. (2012) The principles of designing of algorithm for speech synthesis from texts written in Albanian language. International Journal of Computer Science Issues, Vol. 9, No 3, May 2012.
7. Hastie.T, Tibshirani.R, Friedman.J. (2009). "The Elements of Statistical Learning" Data Mining, Inference, and Prediction Second Edition.WIELY.. pp 119-122.
8. Holmes D. I. (1998). Authorship attribution. Literary and Linguistic Computing, 13(3):111–117.
9. Hosmer.D.V, Lemeshow.S, Sturdivant. R. (2013). Applied Logistic Regression. Third Edition, John Wiley & Sons. pp 1-20, 35-45, 269-278.
10. James.G , Witten.D , Hastie.T ,Tibshirani. R. (2013). "An Introduction to Statistical Learning" with applications in R. Springer pp 127-137.
11. Kennedy E, Levitt H, Neuman AC, Weiss M. (1998). Consonant-vowel intensity ratios for maximizing consonant recognition by hearing-impaired listeners. Journal of the Acoustical Society of America. ;103(2):1098-114.
12. Khmelev, D.V., & Tweedie, F.J. (2001). Using Markov chains for identification of writers. Literary and Linguistic Computing, 16(4), 299–307.
13. Kjell, B. (1994). Authorship determination using letter-pair frequency features with Neural Network classifiers. Literary and Linguistic Computing, 9, 119–124.
14. Kukushkina, O.V., Polikarpov, A.A., & Khmelev, D.V. (2001). Using literal and grammatical statistics for authorship attribution. Problems of Information Transmission, 37(2), 172-184.
15. Mandelbrot, B. (1962). On the theory of word frequencies and on related markovian models of discourse. Structure of language and its mathematical aspects, 190–219.
16. Orr .S, Montgomery.A, Healy.E, Dubno.J. (2010). Effects of consonant-vowel intensity ratio on loudness of monosyllabic words. Journal of the Acoustical Society of America. Nov; 128(5): 3105–3113.
17. Paci.H, Karaj.E, Tafaj.I, Trendafili. E, Salillari.D. (2011) "Author identification in Albanian language" NBIS 2011.

18. Pollo.T, Kessler. B, Treiman. R. (2005). Vowels, syllables, and letter names: Differences between young children's spelling in English and Portuguese. J. Experimental Child Psychology 92 (2005) 161–181
19. Port, R.F. Dalby, J. (1982). Consonant/vowel ratio as a cue for voicing in English. Perception & Psychophysics.
20. Ross M. Sheldon (2007). Introduction to Probability Models. Ninth Edition ELSEVIER 2007 p:185-193
21. Rushiti.R. (2012) Lidhëzat Që dhe Se në gjuhën Shqipe. p: 352-365.
22. Salillari.D, Prifti.L (2016). A multinomial logistic regression model for text in Albanian language, Journal of Advances in Mathematics, Volume 12 Number 07.
23. Salillari.D, Prifti.L (2016). Comparison Study of Logistic Model for Albanian Texts, Journal of Advances in Mathematics, Volume 12 Number 09.
24. Salillari.D, Prifti.L, Kuka.Sh. (2012). "Logistic regression for authorship attribution in albanian text " 7th Annual Meeting of Institute Alb-Shkenca, Conference Of Natural Sciences.
25. Salillari.D, Prifti.L, Shehu.Sh. (2012). "Relative entropy for markov chains of letters in albanian texts", ICEOS 2012.
26. Salillari.D, Prifti.L, Shehu.Sh. (2012). Vleresimi i entropisë në tekstet shqip. Buletini UNIEL Vol 1, 2012, nr 34/10.
27. Sammeth CA, Dorman MF, Stearns CJ. (1999). The role of consonant-vowel amplitude ratio in the recognition of voiceless stop consonants by listeners with hearing impairment. Journal of speech, Language and Hearing Research.
28. Zhao. Y. (2012). R and Data Mining: Examples and Case Studies. First Edition Elsevier.
29. Zipf, George K. (1949). Human behavior and the principle of least effort. Cambridge (MA): Addison-Wesley.