STILOMETRIK TAHLIL VA MASHINAVIY O‘RGANISH METODLARINING LINGVISTIK EKSPERTIZADAGI O‘RNI: XALQARO TADQIQOTLAR SHARHI

Abdurayimova Durdona Samat qizi,
ToshDO‘TAU tayanch doktoranti
abdurayimovadurdona21@gmail.com

DOI: https://doi.org/10.5281/zenodo.17932284

Annotatsiya. *Ushbu maqolada lingvistik ekspertiza sohasida stilometrik tahlil, statistik metodlar va mashinaviy o‘rganish yondashuvlarining qo‘llanishi xalqaro ilmiy adabiyotlar asosida keng qamrovli tarzda sharhlab beriladi. Stilometriya yozma matnlarning leksik, sintaktik, ortografik va paralingvistik belgilarini miqdoriy o‘lchashga asoslangan bo‘lib, mualliflik, yosh, jins yoki platformaga xos xususiyatlarni aniqlashda samarali vosita hisoblanadi. Statistik metodlar — jumladan, ANOVA, MANOVA va t-testlar — stilometrik belgilar o‘rtasidagi farqlarni aniqlash va mashinaviy o‘rganish modellari uchun farqlovchi xususiyatlarni tanlashda muhim ahamiyat kasb etadi. Mashinaviy o‘rganish algoritmlari esa matn tasnifi, shaxs identifikatsiyasi va til xulqini avtomatlashtirilgan tarzda baholashga imkon beradi. Maqolada ushbu uch metodologiyaning xalqaro miqyosda qanday uyg‘unlashtirilayotgani va ularning lingvistik ekspertizadagi roli batafsil yoritiladi, shuningdek, yetakchi tadqiqotlarning natijalari taqqoslanadi.*

Kalit so‘zlar: *stilometrik tahlil, statistik metodlar, mashinaviy o‘rganish, lingvistik ekspertiza, yosh identifikatsiyasi, matn tasnifi, xalqaro tadqiqotlar.*

Annotation.*This article provides a comprehensive review of the application of stylometric analysis, statistical methods, and machine learning approaches in the field of forensic linguistics based on international scholarly literature. Stylometry, which relies on the quantitative measurement of lexical, syntactic, orthographic, and paralinguistic features of written texts, serves as an effective tool for identifying authorship, age, gender, and platform-specific characteristics. Statistical methods—including ANOVA, MANOVA, and t-tests—play a crucial role in detecting differences between stylometric features and selecting discriminative variables for machine learning models. Machine learning algorithms, in turn, enable automated text classification, author identification, and the analysis of linguistic behavior. The article discusses how these three methodological approaches are being integrated at the international level and examines their role in forensic linguistic analysis, while also comparing the findings of leading studies in the field.*

Keywords: *stylometric analysis, statistical methods, machine learning, forensic linguistics, age identification, text classification, international research.*

Raqamli kommunikatsiya vositalarining jadal rivojlanishi natijasida internet tarmoqlarida hosil bo‘layotgan yozma matnlarning hajmi va xilma-xilligi keskin oshdi. Ijtimoiy tarmoqlar, bloglar, xabar almashish platformalari va onlayn forumlar bugungi kunda shaxslararo muloqotning asosiy kanallaridan biriga aylandi. Bunday muhitda shaxsning til xususiyatlari, yozma nutqdagi uslubiy izlari lingvistik ekspertiza va shaxs identifikatsiyasi uchun muhim ma’lumot manbaiga aylanmoqda. Lingvistik ekspertizada shaxsga xos yozuv uslubini aniqlashda bir nechta metodologik yondashuvlar qo‘llaniladi. Ularning eng asosiylari quyidagilardir:

1. Stilometrik tahlil — matnning leksik, sintaktik, ortografik va paralingvistik belgilarini miqdoriy tahlil qilish orqali uslubiy profil yaratish.
2. Statistik metodlar — stilometrik belgilar guruhlar yoki shaxslar o‘rtasida farqlanadimi-yo‘qmi, degan savolga javob beruvchi matematik vositalar (masalan, ANOVA, MANOVA, t-test).
3. Mashinaviy o‘rganish algoritmlari — matnlarni tasniflash va shaxsni avtomatik tarzda identifikatsiya qilish imkonini beruvchi hisoblash yondashuvlari.

Xalqaro tadqiqotlar ko‘rsatishicha, aynan shu uch yo‘nalishning uyg‘unligi shaxsning yoshi, jinsi, ta’lim darajasi yoki muallifligini aniqlashda yuqori aniqlik (80–95 %) ko‘rsatadi. [Koppel, Schler, Argamon, 2009.; Neal va bosh, 2017.; Nguyen va bosh, 2013.]. Masalan, Burrovs klassik stilometrik belgilar asosida adabiy matnlar muallifligini aniqlashga muvaffaq bo‘lgan bo‘lsa[Burrovs, 63–67], zamonaviy tadqiqotlarda bu belgilar mashinaviy o‘rganish modellari bilan birgalikda ishlatilib, real va katta hajmdagi ma’lumotlar ustida qo‘llanmoqda[Stamatatos, 2009.540–550 ; Zhao, Zobel, 2005.5-11].

Statistik metodlar va ularning stilometriyadagi roli. Stilometrik tadqiqotlarda statistik metodlar shaxs yoki guruhlar o‘rtasidagi yozma nutq xususiyatlaridagi farqlarni aniqlash va bu farqlarning ahamiyat darajasini matematik jihatdan isbotlash uchun qo‘llaniladi. Statistik yondashuvlar stilometrik ko‘rsatkichlarning (masalan, o‘rtacha gap uzunligi, leksik boylik, tinish belgilari chastotasi) tasodifiy tebranishlar emas, balki barqaror va farqlovchi belgilar ekanligini aniqlashda muhim vositadir [Field, 2013. 295–310; Johnson ,Wichern, 2007. 389–402].

ANOVA – guruhlararo farqlarni aniqlash. ANOVA (Analysis of Variance) usuli yozma matn ko‘rsatkichlarining turli guruhlar (masalan, yosh, jins, platforma) o‘rtasidagi o‘rtacha qiymatlarida sezilarli farqlar mavjudligini tekshirishga xizmat qiladi. Masalan, yosh guruhlari bo‘yicha o‘rtacha gap uzunligini solishtirishda ANOVA yordamida tilning yoshga xos xususiyatlarini aniqlash mumkin. Klassik stilometrik tadqiqotlarda ANOVA ko‘pincha bir yo‘nalishli shaklda qo‘llanadi. Masalan, Holmes turli mualliflarning matnlarida funksional so‘zlar chastotasini solishtirib, farqlar statistik jihatdan sezilarli ekanini ko‘rsatgan bo‘lsa[Holmes, 1998. 113–115]. Zamonaviy tadqiqotlarda esa bu usul yosh[Nguyen va bosh, 2013. 442–446], jins[Argamon va bosh, 2003. 332–340] va platforma[Stamatatos, 2009. 540–544] bo‘yicha til belgilarini tahlil qilishda keng ishlatiladi. ANOVA natijalarida F-koeffitsiyent, erkinlik darajalari (df) va p-qiymat asosiy statistik ko‘rsatkichlar hisoblanadi. $p < 0.05$ bo‘lganda guruhlar o‘rtasida sezilarli farq bor degan xulosa qilinadi [Field, 2013. 295–310].

MANOVA – ko‘p belgili farqlarni aniqlash. MANOVA (Multivariate Analysis of Variance) usuli bir vaqtning o‘zida bir nechta stilometrik belgilarni (masalan, gap uzunligi, tinish belgilari chastotasi, leksik boylik) guruhlar bo‘yicha taqqoslash imkonini beradi[Johnson, Wichern, 2007. 389–402]. Masalan, o‘smirlar va kattalar yozishmalarida gap uzunligi bilan birga orfografik belgilar (masalan, cho‘zib yozish “zooorrr”) va emoji ishlatish chastotasi ham tahlil qilinadi. MANOVA bu belgilar birgalikda guruhlarni farqlashda qanday rol o‘ynashini aniqlashga yordam beradi. Bu yondashuv mashinaviy o‘rganishdan oldingi bosqichda eng informativ belgilarni tanlash imkonini beradi va yosh guruhlarini stilometrik profil bo‘yicha ajratishda samarali hisoblanadi [Nguyen va bosh, 2013. 442–446; Neal va bosh, 2017.8–14].

T-test – ikki guruhni taqqoslash. T-test ikki guruh orasidagi stilometrik ko‘rsatkichlar farqini tekshirish uchun ishlatiladi[Howell, 2012. 147–180]. Masalan, Facebook va Telegram yozishmalaridagi o‘rtacha gap uzunligini solishtirishda mustaqil t-test qo‘llaniladi. Agar $t(58)=2.45$, $p=0.017$ bo‘lsa, platformalar o‘rtasida sezilarli farq bor degan xulosa chiqariladi. T-test shuningdek, gender farqlarini[Argamon va bosh, 2003. 332–340] yoki stilometrik belgilar bo‘yicha ikki yosh guruhining til xususiyatlaridagi farqlarni aniqlashda keng qo‘llanadi. Bu usul ayniqsa matn korpuslarining kichik guruhlari bo‘yicha sezilarli farqlarni aniqlashda samarali.

Statistik metodlar va mashinaviy o‘rganish bog‘liqligi. Stilometrik tadqiqotlarda statistik testlar ko‘pincha mashinaviy o‘rganish modellari bilan birgalikda qo‘llanadi. Avval stilometrik belgilar statistik jihatdan tekshiriladi (masalan, ANOVA yordamida yosh guruhlari bo‘yicha qaysi belgilar eng farqlovchi ekani aniqlanadi), so‘ng ushbu belgilar mashinaviy o‘rganish modeliga kiritiladi. Nguyen va boshqalar Twitter matnlarida emoji chastotasi, so‘z uzunligi va tinish belgilari bo‘yicha yosh guruhlarini statistik farqlab, keyin SVM modelida tasniflashda 84 % aniqlikka erishgan[Nguyen va boshqalar, 2013. 442–446]. Shu tariqa statistik metodlar xususiyat tanlash (feature selection) bosqichida asosiy rol o‘ynaydi. ANOVA, MANOVA va t-testlar stilometrik tahlilda til birliklarining guruhlar bo‘yicha sezilarli farqlarini matematik asosda ko‘rsatadi. Bu usullar mashinaviy o‘rganishdan avvalgi bosqichda eng farqlovchi belgilarni aniqlash va model natijalarini validatsiya qilish imkonini beradi. Shu sababli zamonaviy xalqaro stilometrik tadqiqotlarda statistik metodlar va mashinaviy o‘rganish birgalikda qo‘llanilishi odatiy holga aylangan [Stamatatos, 2009. 540–544; Neal, 2017. 8–14].

Mashinaviy o‘rganish yondashuvlari. Stilometrik tahlil natijasida olingan belgilarni avtomatik tarzda tahlil qilish va tasniflash uchun mashinaviy o‘rganish (machine learning, ML) algoritmlari keng qo‘llaniladi. Bunday algoritmlar yozma

nutqning turli stilometrik xususiyatlari asosida shaxsning muallifligi, yoshi, jinsi yoki platforma xususiyatlarini yuqori aniqlik bilan (80–95 %) ajrata oladi [Koppel va bosh, 2009. 14–20; Neal va bosh, 2017. 8–14]. ML yondashuvlarining afzalligi shundaki, ular katta hajmdagi matn korpuslarini qayta ishlay oladi, statistik bog‘liqliklarni aniqlaydi va stilometrik belgilar o‘rtasidagi murakkab o‘zaro munosabatlarni modellaydi.

Quyida xalqaro stilometrik tadqiqotlarda keng qo‘llaniladigan asosiy algoritmlar tahlil qilinadi. Naive Bayes klassifikatori stilometrik tahlilda samarali usullardan biridir. U Bayes teoremasiga asoslanadi va belgilar o‘zaro mustaqil deb faraz qiladi. Ushbu algoritm funksional so‘zlar chastotasi yoki n-gram modellari asosida mualliflikni aniqlashda samarali natijalar beradi[Zhao, Zobel, 2005. 5–11]. Koppel va Schler bir muallifga tegishli matnlarni tasdiqlash (authorship verification) muammosini Naive Bayes usuli yordamida hal qilib, yuqori aniqlikka erishgan[Koppel va Schler, 2004. 490–492]. Ushbu yondashuv, ayniqsa, korpus hajmi cheklangan bo‘lgan hollarda samarali bo‘lib, ko‘pincha bazaviy model sifatida ishlatiladi.

Support Vector Machines (SVM). algoritmlari stilometrik tahlilda keng qo‘llanadi, chunki ular yuqori o‘lchamli belgilar makonida sinflararo eng yaxshi chegarani aniqlash imkonini beradi[Joachims, 1998. 138–140] Nguyen va boshqalar Twitter foydalanuvchilarining yoshini aniqlash bo‘yicha tadqiqotida SVM yordamida 80–84 % aniqlikka erishgan. Ular belgilar sifatida emoji chastotasi, tinish belgilaridan foydalanish, cho‘zib yozishlar va leksik birliklar statistik taqsimotidan foydalangan[Nguyen va boshqalar, 2013. 442–446]. SVM algoritmi ko‘pincha ANOVA yoki MANOVA yordamida farqlovchi belgilar tanlangach qo‘llanadi — bu esa modelning aniqligini oshiradi.

Random Forest algoritmi ko‘plab qaror daraxtlarining kombinatsiyasiga asoslanadi va stilometrik tahlilda katta hajmdagi, murakkab belgilar to‘plami bilan ishlashda samarali [Stamatatos, 2009. 540–544] Neal va boshqalar Random Forest yordamida blog yozuvlari muallifligini aniqlash bo‘yicha tajribalar o‘tkazib, model barqarorligi va yuqori aniqligini ta’kidlashgan[Neal va bosh, 2017. 8–14]. Ushbu usul ayniqsa belgilar soni ko‘p, lekin ularning ayrimlari ortiqcha bo‘lishi mumkin bo‘lgan holatlarda foydali, chunki u xususiyatlarning ahamiyatini avtomatik baholaydi.

Linear Discriminant Analysis (LDA) algoritmi mashinaviy o‘rganishning klassik usullaridan biri bo‘lib, belgilar o‘lchamlarini kamaytirish va guruhlararo farqlarni aniqlashda ishlatiladi[Burrows, 1987. 63–67; Johnson , Wichern, 2007. 389–402]. Ushbu yondashuv ko‘pincha yosh guruhlarini yoki jinsga oid farqlarni stilometrik belgilar bo‘yicha tasniflashda qo‘llaniladi. Argamon va boshqalar jins

bo‘yicha yozma matn uslubini ajratishda diskriminant tahlil yordamida yuqori tasniflash natijalariga erishgan[Argamon va boshqalar, 2003. 332–340]. LDA ko‘pincha statistik tekshiruvlar (ANOVA, MANOVA) bilan birga ishlatiladi, chunki u guruhlararo farqlarni maksimal ajratuvchi kombinatsiyalarni topadi.

Xalqaro tajribada mashinaviy o‘rganish algoritmlari ko‘pincha stilometrik tahlil va statistik tekshiruvlar bilan uch bosqichli tizim sifatida integratsiyalashgan:
- stilometrik belgilar ajratiladi (masalan, gap uzunligi, n-gramlar, emoji chastotasi);
- statistik metodlar yordamida (ANOVA, MANOVA, t-test) farqlovchi belgilar tanlanadi;
- tanlangan belgilar mashinaviy o‘rganish modeliga (SVM, Random Forest, LDA va b.) beriladi. Masalan, Nguyen va boshqalar Twitter matnlarida yosh guruhlari bo‘yicha statistik farqlarni aniqlab, SVM modeli orqali yuqori aniqlikka erishgan[Nguyen va bosh, 2013. 442–446]. Neal va boshqalar esa stilometrik belgilarni Random Forest yordamida barqaror tasniflash tizimiga joylashtirgan[Neal, 2017. 8–14]. Bunday yondashuv stilometrik ekspertizani avtomatlashtirishda va katta korpuslarda yuqori aniqlik ko‘rsatkichlariga erishishda hal qiluvchi rol o‘ynaydi[Koppel, 2009.14–20; Stamatatos, 2009.540–544]. Mashinaviy o‘rganish yondashuvlari stilometrik tahlilda shaxsga xos yozuv uslubini aniqlash, yosh va jins kabi sotsiolingvistik belgilarni tasniflash, shuningdek, mualliflikni aniqlashda keng imkoniyatlar yaratmoqda. Naive Bayes, SVM, Random Forest va LDA algoritmlari xalqaro tadqiqotlarda eng ko‘p qo‘llaniladi. Ular statistik metodlar bilan uyg‘unlashganda yuqori aniqlik va ishonchlilikka erishadi. Shu bois bugungi kunda lingvistik ekspertiza sohasida stilometriya va mashinaviy o‘rganishning integratsiyasi muhim ilmiy yo‘nalish sifatida shakllangan.

Neyron tarmoqlar va chuqur o‘rganish (Deep Learning). So‘nggi yillarda stilometrik tahlilda chuqur o‘rganish modellari (Deep Learning) keng qo‘llanila boshladi. Ayniqsa RNN (Recurrent Neural Networks), CNN (Convolutional Neural Networks) va Transformer arxitekturasi asosidagi modellarning matn tasnifida samaradorligi yuqori[Neal, 2017. 20–26]. Masalan, BERT (Bidirectional Encoder Representations from Transformers) modeli matn kontekstini chuqur tahlil qilish imkonini beradi. Bu yondashuv stilometrik belgilarni oldindan belgilamasdan, modelning o‘zi ularni ichki qatlamlarda o‘rganishini ta’minlaydi. Shuning uchun chuqur o‘rganish usullari katta korpuslar bo‘lgan holatlarda, masalan, Twitter yoki bloglar tahlilida, klassik SVM yoki Naive Bayesdan ancha yuqori aniqlikka erishmoqda[Nguyen, 2020. 55–61].

Ensemble modellari va gibrid yondashuvlar. Ensemble modellari bir nechta klassifikatorlarni birlashtirish orqali natijani yaxshilashga qaratilgan. Masalan,

Random Forest + SVM kombinatsiyasi yoki Naive Bayes + LDA ketma-ketligi. Koppel va Ordóñez blog muallifligini aniqlashda bir nechta algoritmlarni gibrid shaklda qo‘llab, bitta modelga qaraganda 6–8 % yuqori aniqlikka erishgan[Koppel, Ordóñez, 2011. 127–133]. Gibrid yondashuvlar, ayniqsa, stilometrik belgilar murakkab bo‘lgan holatlarda (masalan, emoji, tinish belgisi, kod-switching, grafostilistik elementlar birgalikda mavjud bo‘lsa) foydali bo‘ladi.

Feature engineering va model barqarorligi. Mashinaviy o‘rganishda model aniqligini oshirish faqat algoritm tanlashga emas, balki belgilarni to‘g‘ri tanlash (feature engineering) jarayoniga ham bog‘liq[Hair, 2010. 112–135]. Masalan, gap uzunligi, leksik boylik, morfologik markerlar, emoji chastotasi, transliteratsiya darajasi, cho‘zib yozishlar kabi belgilar birgalikda olinishi tasnif modelining barqarorligini oshiradi. Stamatatos shuni ta’kidlaydiki, belgilarni kombinatsiyalab ishlatish SVM va Random Forest modellari samaradorligini sezilarli darajada oshiradi[Stamatatos, 2009. 544–550].

Platformaga xos model farqlari. Yana bir muhim jihat — mashinaviy o‘rganish modellari platformaga (Facebook, Telegram, Twitter, bloglar) qarab farq qiladi. Masalan, Twitter matnlari qisqa, lekin emoji, hashtag, kod-switching elementlariga boy; Facebook matnlari esa ko‘proq sintaktik murakkablikka ega. Shuningdek, Twitterdagi yoshni aniqlash modellari Facebookga bevosita o‘tkazilganda aniqlik 10–15 % ga tushib ketadi[Nguyen, 2013. 442–446]. Shu sababli zamonaviy tadqiqotlarda platformaga mos maxsus modellash (platform-specific modeling) yondashuvi shakllangan.

Model natijalarini baholash mezonlari. Mashinaviy o‘rganish modellari samaradorligi faqat aniqlik (accuracy) ko‘rsatkichi bilan emas, balki F1-score, precision, recall, ROC-AUC kabi ko‘rsatkichlar bilan baholanadi. Bu ayniqsa nomutanosib guruhlar bo‘lgan holatlarda, masalan, 13–19 yoshdagi va 50+ foydalanuvchilar soni farq qilsa. Yoshni tasniflash bo‘yicha tajribalarida Random Forest modelining F1-score ko‘rsatkichi 0.82 bo‘lganini, bu esa modelning muvozanatli ishlashini ko‘rsatadi[Nguyen, 2020. 57–59]. Mashinaviy o‘rganish sohasida klassik modellardan (Naive Bayes, SVM, Random Forest, LDA) tashqari chuqur o‘rganish, ensemble va platformaga moslashgan yondashuvlar tobora keng qo‘llanmoqda. Bular stilometrik tahlilning aniqligi, barqarorligi va amaliy ahamiyatini oshiradi. Xususan, yosh, jins, mualliflik va kommunikativ platformani avtomatik aniqlash bo‘yicha xalqaro tadqiqotlarda bu modellarning uyg‘unligi yuqori samaradorlik bermoqda.

Stilometrik tahlil, statistik metodlar va mashinaviy o‘rganish yondashuvlari mualliflik ekspertizasida muhim hisoblanadi. Bu metodlar xalqaro miqyosda

muallifning demografik belgilarini aniqlashda keng qo‘llaniladi. Ayrim tadqiqotlar klassik lingvistik belgilarga(funksional so‘zlar, gap uzunligi) asoslanib statistik testlardan foydalanib mualliflikni aniqlashga qaratilgan bo‘lsa, keyingi tadqiqotlarda mashinaviy o‘rganish algoritmlari (Naive Bayes, SVM, Random Forest) keng joriy etildi. Shuningdek, so‘nggi yillarda esa chuqur o‘rganish modellarining (RNN, CNN, BERT) qo‘llanilishi stilometrik tahlilning aniqligini sezilarli oshirdi.

Quyida ushbu yo‘nalishdagi 10 ta asosiy xalqaro tadqiqotning platformasi, stilometrik belgilar turi, statistik va mashinaviy yondashuvlari, shuningdek aniqlik darajalari bo‘yicha tizimli taqqoslanmasi keltirilgan.

4.1-jadval. Stilometrik tahlil va mashinaviy o‘rganish bo‘yicha xalqaro tadqiqotlar taqqoslash

T/r	Muallif, yil	Platforma, ma’lumot manbasi	Stilometrik belgiar	Statistik metodlar	ML model	Tadqiqot maqsadi	Aniqlik
1	Mosteller va Wallace (1964)	Federalist Papers (ingliz matnlari)	Funksional so‘zlar	chastotasi t-test	-	Mualliflikni aniqlash	87 %
2	Burrovs (1987)	Adabiy romanlar	Gap uzunligi, funksional so‘zlar	ANOVA	LDA	Mualliflik va uslub farqlari	82%
3	Argamon va boshqa (2003)	Akademik matnlar	Leksik, sintaktik, funksional belgilar	MANOVA, t-test	LDA	Jinsga oid farqlar	76 %
4	Koppel va Schler(2004)	Blog yozuvlari	***N***-gramla, funksional belgilar	—	Naive Bayes	Mualliflikni tekshirish (verification)	88 %
5	Zhao va Zobel (2005)	Bloglar va forumlar	Funksional so‘zlar chastotasi	*T-test*	Naive Bayes	Mualliflikni aniqlash	85 %
6	Stamatato (2009)	Blog, maqola, arxiv matnlari	Belgilar kombinatsiyasi	ANOVA SVM	RF	Mualliflik atributsiyasi	90 %
7	Nguyen va boshqalar (2013)	Twitter	Emoji chastotasi, tinish belgilari, cho‘zib yozishlar	ANOVA SVM		Yosh bo‘yicha tasniflash	80–84 %

8	Koppel va Ordóñez, (2011)	*Bloglar*	Belgilar kombinatsiyasi	—	Ensemble	Mualliflik aniqligi oshirish	92 %
9	Neal va boshqalar (2017)	Blog va forumlar	Stilometrik belgilar to‘plami	MANOVA	RF, Deep Learning	Mualliflik va uslub profilini aniqlash	88–93 %
10	Nguyen va boshqalar (2020)	Twitter, Facebook	Kontekstual belgilar (BERT)	—	Deep Learning	Yosh va jins profilingi	90–94 %

Jadval xronologik tartibda (1964–2020) tuzilgan bo‘lib, stilometriya sohasining tarixiy rivojlanish bosqichlarini yaqqol aks ettiradi. Birinchidan, platformalar turlicha: Federalist Papers, adabiy romanlar, blog yozuvlari, forumlar, Twitter va Facebook. Shuningdek, stilometrik belgilar tarkibi vaqt o‘tishi bilan murakkablashgan: funksional so‘zlardan tortib, emoji va kontekstual embeddinglargacha hamda statistik metodlardan klassik testlar (t-test, ANOVA, MANOVA), mashinaviy modellardan Naive Bayes, SVM, RF, chuqur o‘rganish usullari qo‘llangan va aniqlik ko‘rsatkichlari zamonaviy yondashuvlarda 90 % dan yuqori darajaga yetgan. Umuman olganda yillar o‘tishi bilan aniqlikni oshirish uchun tadqiqot metodlari va yondashuvlari murakkablashib borgan. Ushbu bosqichlar ketma-ketligi lingvistik ekspertiza uchun metodologik asos yaratadi va shaxsni yosh, jins yoki mualliflik belgilariga ko‘ra aniqlashda yuqori natijalar beradi.

Xulosa qilib aytganda, stilometrik tahlil, statistik metodlar va mashinaviy o‘rganish texnologiyalarining uyg‘unligi hozirgi zamon lingvistik ekspertizasi uchun eng ilg‘or yondashuvlardan biri sifatida shakllandi. Dastlabki tadqiqotlar leksik va sintaktik ko‘rsatkichlarni statistik jihatdan tahlil qilish orqali mualliflikni aniqlashga qaratilgan bo‘lsa, 2000-yillardan boshlab mashinaviy o‘rganish algoritmlari bu jarayonni avtomatlashtirishga keng yo‘l ochdi. Statistik metodlar — jumladan ANOVA, MANOVA va t-testlar — stilometrik belgilar o‘rtasidagi farqlarni matematik asoslash va mashinaviy o‘rganish modellari uchun eng farqlovchi belgilarni tanlashda hal qiluvchi rol o‘ynaydi. Ayniqsa, yosh, jins va platformaga oid til xulqini farqlashda bu metodlar ishonchli asos yaratadi. Mashinaviy o‘rganish yondashuvlari (Naive Bayes, SVM, Random Forest, LDA) va chuqur o‘rganish modellarining (BERT, RNN, CNN) joriy etilishi stilometrik tahlilning aniqligini sezilarli darajada oshirdi. Xalqaro tadqiqotlar shuni ko‘rsatadiki, stilometrik belgilarni statistik tekshiruvlar bilan boyitish va ularni mashinaviy o‘rganish modellari bilan birlashtirish orqali matn egasining yoshi, jinsi yoki muallifligini 80–95 % aniqlikda aniqlash mumkin[Nguyen,2013, 2017, 2020]. Bundan tashqari,

platformaga xos modellash (Twitter, Facebook, bloglar), ansambl va gibrid yondashuvlar, shuningdek feature engineering usullarining qo‘llanilishi stilometrik ekspertizani yanada kuchaytirmoqda. Ayniqsa, chuqur o‘rganish usullari katta hajmdagi ijtimoiy tarmoq matnlarini kontekstual darajada qayta ishlash imkonini bermoqda, bu esa klassik modellarga nisbatan yuqori aniqlikni ta’minlaydi. Xalqaro tajribani umumlashtirgan holda aytish mumkinki:

1. Stilometrik tahlil yozma nutqdagi leksik, sintaktik va ortografik belgilarni tizimli miqdoriy tahlil qilish imkonini beradi.
2. Statistik metodlar belgilar o‘rtasidagi farqlarni matematik jihatdan asoslashga xizmat qiladi.
3. Mashinaviy o‘rganish yondashuvlari bu belgilar asosida matn egasi haqida avtomatik tarzda yuqori aniqlikda xulosa chiqarishni ta’minlaydi.

Ushbu uch metodologiyaning integratsiyasi lingvistik ekspertiza sohasida shaxsni identifikatsiya qilish, yosh va jins profilingini aniqlash, platformaga xos til xulqini o‘rganish, shuningdek, raqamli xavfsizlik va sud-lingvistik ekspertizalarda katta imkoniyatlar yaratmoqda. O‘zbek tili materiallari bo‘yicha ham stilometrik ko‘rsatkichlarni statistik va mashinaviy metodlar bilan uyg‘unlashtirish istiqbolda til xulqini yoshga qarab farqlashni ilmiy asosda aniqlash imkonini beradi.

Foydalanilgan adabiyotlar:

1. Mosteller F., Wallace D. Inference and Disputed Authorship: The Federalist. — Reading: Addison-Wesley, 1964. – P. 48–52.
2. Burrows J. Word-patterns and story-shapes: The statistical analysis of narrative style // Literary and Linguistic Computing. – 1987. – Vol. 2. – №2. – P. 63–67.
3. Argamon S., Koppel M., Fine J., Shimoni A. Gender, genre, and writing style in formal written texts // Text. – 2003. – Vol. 23(3). – P. 332–340.
4. Koppel M., Schler J. Exploiting stylistic idiosyncrasies for authorship attribution // Computers and the Humanities. – 2004. – Vol. 38. – P. 490–492.
5. Zhao Y., Zobel J. Effective and scalable authorship attribution using function words // Information Retrieval. – 2005. – Vol. 8. – P. 5–11.
6. Stamatatos E. A survey of modern authorship attribution methods // Journal of the American Society for Information Science and Technology. – 2009. – Vol. 60(3). – P. 540–550.
7. Koppel M., Ordóñez C. Ensemble methods in authorship attribution // JASIST. – 2011. – Vol. 62(1). – P. 127–133.
8. Nguyen D., Smith N. A., Rosé C. P. Author age prediction from text using linear regression // Proceedings of the 51st ACL. – 2013. – P. 442–446.
9. Neal T. M. S. et al. Surveying stylometry techniques and applications // ACM Computing Surveys. – 2017. – Vol. 50(6). – P. 8–26.
10. Nguyen D., Gravel R., Trieschnigg D., Meder T. "How old do you think I am?": A study of language and age in Twitter // LREC. – 2020. – P. 55–61.
11. Joachims T. Text categorization with support vector machines: Learning with many relevant features // Proceedings of ECML. – 1998. – P. 137–142.

12. Hair J. F., Black W. C., Babin B. J., Anderson R. E. Multivariate Data Analysis. – 7th ed. – Pearson, 2010. – P. 112–135.
13. Field A. Discovering Statistics Using IBM SPSS Statistics. – London: SAGE, 2013. – P. 247–263.
14. Neal T. M. S. et al. Modern stylometry techniques for authorship attribution in large social media corpora // Digital Scholarship in the Humanities. – 2018. – Vol. 33(3). – P. 467–482.
15. Schler J., Koppel M., Argamon S., Pennebaker J. Effects of age and gender on blogging // AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs. – 2006. – P. 199–205.
16. Ng A. Y. Feature selection, L1 vs. L2 regularization, and rotational invariance // Proceedings of ICML. – 2004. – P. 78–85.