UDC: 81’23:004.92 Info M str. 26-30

УТВРЂИВАЊЕ ИДЕНТИТЕТА АУТОРА НА ОСНОВУ АНАЛИЗЕ ТЕКСТА
DETERMINING THE IDENTITY OF THE AUTHOR BASED ON THE ANALYSIS OF THE TEXT

Marko Gogić, Fakultet organizacionih nauka, Univerzitet u Beogradu

РЕЗИМЕ: За разлику од „обичних“, сајбер криминалци не морају да брину да ће случајно оставити отиске прстију или трагове ДНК који би одали њихов идентитет. Уместо тога они могу слободно да се крећу кроз сајбер простор и користе друштвене мреже, онлајн блогове или мејлове за размену илегалног материјала, слање претећих порука или „фишинг“, кријӯћи се притом иза својих дигиталних, псеудо-идентитета које могу лако да промене. Али оно што не може лако да се промени јесу индивидуалне језичке карактеристике које попут отисака прстију остају утиснуте у електронском текстуалном материјалу. Управо јединствена језичка обележја појединца представљају биометријску грађу на основу које форензичка лингвистика може да утврдити идентитет аутора неке претеће поруке или захтева за откуп. Метода форензичке лингвистике која се користи у ту сврху назива се стилометрија. Обиље доступног електронског текстуалног материјала, уз одсуство класичних биометријских трагова у сајбер простору, намеће потребу за све већом применом стилометрије ради решавања низа практичних проблема. Циљ овог рада је да испита валидност постојећих стилометријских решења за идентификацију аутора кроз анализу текста заснованих на примени алгоритама машинског учења. У ту сврху у раду ће бити приказан историјски развој стилометрије, уз анализу и критички осврт на постојеће радове који приказују најбоља решења из ове области.

КЉУЧНЕ РЕЧИ: Форензичка лингвистика; стилометрија; машинско учење; лингвистички маркери; атрибуција ауторства.

ABSTRACT: Unlike the „ordinary“ criminals, cyber criminals do not have to worry that they will accidentally leave fingerprints or traces of DNA that would reveal their identity. Instead, they can move freely through cyberspace and use social networks, online blogs for exchange of illegal materials, sending threatening messages or „phishing“, hiding behind their digital, pseudo-identities that can easily be changed. However, what cannot be easily changed is the individual language characteristics that, just like the fingerprints, remain embedded in electronic textual material. Those linguistic features of an individual represent biometric material based on which the forensic linguistics can determine the identity of an author of a threat message or a ransom request. The forensic linguistic method used for this purpose is referred to as the stylometry. The abundance of available electronic textual material, coupled with the absence of classical biometric traces in cyberspace, imposes the need for the increased use of stylometry for solving a number of practical problems. The aim of this paper is to examine the validity of existing stylometric techniques for authorship attribution based on the use of machine learning algorithms. For this purpose, the paper will present a historical overview of the development of the stylometry, along with an analysis and a critical review of the existing work that show the best solutions in this field.

KEY WORDS: forensic linguistics; stylometry; machine learning; linguistic features; authorship attribution.

1. УВОД

Свака особа поседује јединствен начин коришћења језика, било у усменој било писаној комуникацији, где се могу уочити нечија индивидуална својства у погледу говорног израза и употребе језичких средстава на свим нивоима језичке структуре. Појава личних говорних карактеристика и лични стил употребе језичких средстава главни су извори форензичких маркера приликом утврђивања идентитета говорника односно писца, верификовања идентитета говорника (писца), као и приликом утврђивања лингвистичког профила непознатог говорника (писца) [2].

Област форензике која се бави проучавањем употребе личног стила језичких средстава назива се форензичка лингвистика. Форензичка лингвистика у прецизном именовању није наука него вештина примене научних сазнања, теорија и метода лингвистике приликом анализе узорака говорног и писаног језика у истражним поступцима и судским споровима у којима је језик део доказа, а понекад и једини доказ. Њену суштину чини анализа варијантних језичких елемената, односно утврђивање индивидуалних варијација у широком варијационом пољу у неком од појавних облика језика [2].

Форензичка лингвистика се у кривичном поступку може користити за утврђивање идентитета (приписивање ауторства) аутора одређеног текста, верификацију идентитета, профилисање аутора, откривање плагијата, анализу исказа, анализу дискурса, анализу признања и друго. Оно што је фокус интересовања овог рада је примена форензичке лингвистике на пољу утврђивања идентитета аутора кроз анализу текста, односно утврђивање идентитета писца.[1] Метода форензичке лингвистике које се тиме бави назива се стилометрија.

Стилометрија би укратко могла да се дефинише као проучавање јединственог лингвистичког стила и јединствених лингвистичких карактеристика у писаном језику са циљем утврђивања идентитета аутора, односно атрибуције ауторства [4]. Она се заснива на претпоставци да сваки

[1] За више детаља о начинима утврђивања идентитета говорника погледати [25].

појединац има одређене навике у писању специфичне само за њега, као што су основни вокабулар који користи, структура реченица и фразеологија. Другим речима, сматра се да не постоје два аутора која пишу на идентичан начин [8]. Индивидуалне лингвистичке варијације карактеристичне за неког појединца независне су од његове воље, због чега се њима не може свесно манипулисати, што их чини идеалним извором података за примену стилометрије. Стилометрија као метода форензичке лингвистике не настоји само да утврди лингвистичке варијације, већ и статистичке методе неопходне за њихово мерење.

Важно је напоменути да је утрвђивање идентитета на основу обележја језика само по себи сложен задатак јер је човеков језик симултано одређен на биолошком, социолошком и психолошком плану. Сложеност планова на којима језик одређује човека имплицира потребу да се у лингвистичкој форензичкој анализи тражи више удружених индивидуалних маркера да би се идентификација потврдила или оспорила [2].

Идеја да се идентитет неке особе може утврдити помоћу анализе писаног језика није нова и прво је почела да се користи у оквиру књижевне теорије са циљем утврђивања спорних ауторстава књижевних дела. Свакако најпознатија дилема те врсте у књижевности је чувено „Шекспирово питање“, односно питање да ли је стварно Вилјем Шекспир писац великих дела издатих под његовим именом?[2]

Небројано много књига и радова написано је на ту тему, чини се без већег успеха да се дође до коначног одговора. Читава дебата сводила се на покушаје да се докаже да је неко други прави аутор дела приписаних Шекспиру кроз изношење низа аргумената који су се најчешће могли тумачити на различите начине. То је подстакло Томаса Менденхала (1841-1924), америчког физичара и професора са Универзитета у Охају, да напише један од првих радова из области стилометрије (1887). Његова замисао била је да пронађе математички модел заснован на фреквенцији дужине речи, помоћу ког би било могуће утврдити ко је заиста написао неко дело [5]. Иако Менденхал није успео да реши „Шекспирово питање“, његов допринос на пољу квантитативне анализе стила писања био је огроман [6]. Овај рад праћен је радовима Јула (*Yule,* 1938) и Зипфа (*Zipf*, 1932) који су такође применом статистичких метода покушавали да реше питања спорних ауторстава у књижевности [9] [10].

Први већи успех на плану идентификације аутора путем анализа текста остварили су Молестер и Валас (*Mosteller, Wallace*) 1964. године, успевши да реше питање ауторства 12 политичких есеја из 18. века.[3] Њихов рад означио је почетак нетрадиционалних истраживања на пољу стилометрије, која су у односу на традиционална искорачила из искључивог домена књижевности и која су настојала да утврде језичке карактеристике за мерење и квантификовање стила писања [7]. Идентификовање аутора кроз анализу текста је тако пронашло примену и у бројним другим областима као што су: приписивање порука познатим терористима [11]; утврђивање идентитета аутора претећих порука или захтева за откуп, верификација самоубилачких порука, питање ауторских права [12]; идентификација аутора неког злонамерног кода или софтвера [13];

2. ОПИС ПРОБЛЕМА

Утврђивање идентитета аутора кроз анализу текста треба да пружи одговор на питање ко је написао одређени текст, или прецизније речено, ко је из групе кандидата аутор спорног текст.

Карол Часки (*Carole Chaski*), једна од најпознатијих америчких форензичких лингвисткиња и водећи експерт на пољу атрибуције ауторства, наводи бројне случајеве из праксе у којима је било неопходно утврдити идентитет аутора кроз анализу текста. Ти примери на најбољи начин илуструју суштину проблема који се решавају помоћу стилометрије [1].

Случај број 1 - Запослени у државној служби отпуштен је након што је свом надређеном послао поруку увредљиве садржине. Он је након тога тужио државу тврдећи да је било ко од колега могао да приступи његовом компјутеру и пошаље ту поруку без његовог знања.

Случај број 2 - Млад и здрав човек пронађен је мртав у свом стану. Тело је пронашао цимер који је обавестио полицију. Аутопсија је утврдила да је узрок смрти инјекција коју је младић примио, што је полицију навело да посумња да је у питању самоубиство. Током истраге, цимер предаје полицији опроштајне поруке које је пронашао у њиховом заједничком рачунару. Поруке никада нису одштампане, нити су пронађене пре смрти [1].

И у једном и у другом сценарију намеће се логично питање – ко је заиста написао поруке, односно „ко је био за тастатуром“ [1]? У првом случају то може бити било ко од запослених; у другом случају то може бити или жртва или цимер који је пронашао тело.

Елементи проблема су следећи:

1. Постоји текста чији је аутор непознат
2. Постоји круг кандидата за које се сумња да су написали дати текст (најмање два, највише *m* кандидата)
3. Доступни су узорци текстова које је сваки од кандидата (осумњичених) написао

Стилометрија као форензичка метода настоји да постављени проблем реши путем анализе стила писања кандидата (осумњичених). Утврђивање идентитета аутора кроз анализу текста заправо подразумева повезивање једног од аутора из групе са спорним текстом.

3. МЕТОДОЛОГИЈА

Процес утврђивања идентитета аутора кроз анализу текста састоји се из четири итеративне фазе.

[2] Они који би на постављено питање одговорили одрично тврде да је у питању завера енглеске тајне службе и да је драму „Ромео и Јулија“ заправо написао неко други, при чему је листа предложених кандидата подужа.

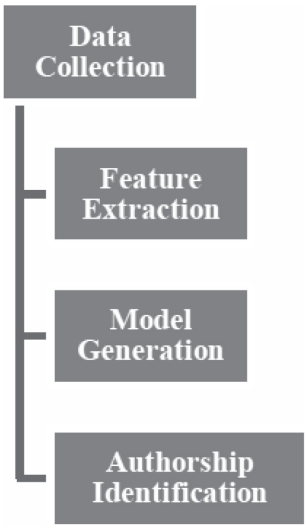
[3] У питању је било 12 спорних есеја из серије политичких текстова тројце аутора познатих као Федералистички списи (*The Federalist Papers*).

У првој фази прикупља се текстуални материјал за који је познатно да је написан од стране кандидата (осумњичених) за ауторство. То могу бити мејлови, поруке, постови са блогова или друштвених мрежа, текстуална документа различите садржине, па и читаве књиге. Карактеристично за ову фазу је и иницијално испитивање прикупљеног материјала које има за циљ да се истраживач упозна са садржајем, тематиком, количином и обимом прикупљеног материјала.

Након иницијалног испитивања приступа се екстракцији лингвистичких макера који чине јединствени отисак за идентификацију аутора. Лингвистички маркери могу бити лексички (дужина реченица, разноврсност вокабулара, најчешће коришћене речи, итд), синтаксни (произилазе из начина на који се организује реченица), структурални (начин на који аутор организује текст, дужина пасуса, увлачење текста, коришћење потписа), маркери везани за сам садржај текста (кључне речи из неке области) и идиосинкратични [19, 20]. Битно је истаћи да у теорији и пракси не постоји сагласност око тога који скуп лингвистичких маркере је најбоље мерити.[4] Такође, квалитет екстракције маркера зависи и од количине доступног материјала за сваког аутора. Према постојећим стандардима сматра се да је неопходно минимум 200 до 250 речи по аутору за успешну екстракцију лингвистичких маркера [8].

Следећи корак представља изградњу модела предвиђања помоћу неког од алгоритама класификације, с обзиром на чињеницу да тврђивање идентитета аутора кроз анализу текста спада у проблеме надгледане мулти-таргет класификације. Маркери екстраковани у претходној фази користи се као улазни атрибути за тренирање модела.

Последњи корак подразумева примену модела и идентификацију аутора уз анализу добијених резултата. Идентификација се врши тако што алгоритам спорни текст сврстава у једну од задатих класа које у овом случају представљају кандидате за ауторство (осумњичене).



Слика 1 – *шематски приказ методологије идентификације аутора кроз анализу текста* ***[24]***

[4] Рудман (*Rudman*) је још 1998. израчунао да је предложено око 1000 различитих сетова маркера [15].

4. ПРЕГЛЕД АЛГОРИТАМА

Најчешће коришћени алгоритми машинског учења за решавање питања атрибуције ауторства су *SVM* и неуронске мреже - *CNN* и *RNN* алгоритми.

SVM (енг. *support vector machines*) је алгоритам надгледаног учења који се користи за решавања проблема класификације и регресије. Изузетно је популаран у студијама које се баве питањем атрибуције ауторства јер се показало да помоћу њега може постићи висок ниво прецизности идентификације. *SVM* проблем класификације решава тако што све маркере (улазне податке) које добије из тренинг сета дели у вишедимензионалном простору помоћу хипер равни или граничних линија. Алгоритам затим проучава оне инстанце података које су веома близу границе са другим категоријама (такозване екстремне случајеве) и на основу тога одлучује у коју категорију да сврста нове инстанце података [21].

Друго популарно решење представљају алгоритми из групе неуронских мрежа. Неуронске мреже су алгоритми засновани на моделу људског мозга, развијени да препознају имплицитну организацију и скривене образце у подацима. Како се стилометрија у основи заснива на препознавању образаца у тексту, примена ових алгоритама у области атрибуције ауторства је честа. Најчешће коришћени алгоритми из ове групе су *CNN* и *RNN*.

5. КРИТИЧКИ ОСВРТ НА ПОСТОЈЕЋА РЕШЕЊА

У обиљу постојећих студија из области стилометрије и атрибуције ауторства овде ће бити анализиране само оне које су постигле висок ниво прецизности идентификације и које су по техничким решењима међусобно упоредиве, уз осврт на њихове карактеристике и слабости. Фактори који отежавају поређење студија су следећи: различити скупови података на којима су спроводена истраживања; различити скупови лингвистичких маркера коришћени за тренирање модела; и различити алгоритми машинског учења коришћени за изградњу модела.

Када су у питању подаци на којима се реализују истраживања, истраживачи могу или да користе неке од јавно доступних тренинг скупови података или да их сами креирају помоћу доступног електронског текстуалног материјала (постови са форума, новински чланци, мејлови, итд). Експериментални сетови података састоје се из скупа докумената са познатим и скупа докумената са спорним ауторством, и у великој мери се разликују у погледу броја аутора, количини и величини текстова за сваког од њих и тематици садржаној у текстовима. Све наведено има значајан утицај на исход предвиђања.

Већ је истакнуто како не постоји сагласност око најбољег скупа лингвистичких маркера које треба мерити, због чега се у пракси користе бројна решења. То има за последицу да чак и исти алгоритми коришћени на истом скупу података дају различите резултате. Ситуација се додатно компликује

студија	скуп података	број аутора/број текстова по аутору	језик	маркери	алгоритам	прецизност мерења
[8]	*ReutersC50*	50/50	енглески	фреквенција речи	*SVM*	85%
[16]	*ReutersC50*	50/50	енглески	фреквенција речи	*SVM*	88%
[17]	*ReutersC50*	50/50	енглески	скуп различитих лингвистичких маркера	*SVM*	91,7%
[22]	*English Internet newsgroup messages*	20/48	енглески	скуп различитих лингвистичких маркера	*SVM*	97.69%
[22]	*Chinese Bulletin Board System (BBS) messages*	20/37	кинески	скуп различитих лингвистичких маркера	*SVM*	88.33%
[11]	серија постова скинутих са интернет форума	20/20	арапски	скуп различитих лингвистичких маркера	*SVM*	94.83%
[11]	серија постова скинутих са интернет форума	20/20	енглески	скуп различитих лингвистичких маркера	*SVM*	97%.

Табела 1 - *приказ анализираних студија*

чињеницом да истраживачима на располагању стоји и велики број различитих алгоритама. Овде ће бити приказана решења заснована на употреби *SVM* алгоритма.

У прве три студиј *SVM* алгоритам коришћен је на истом скупу података[5] уз добијене различите резултате. Резултати од 85% и 88% постигнути су уз мерење фреквенције речи (лексички маркери) као главних језичких обележја [8] [16]. Успешност од 91,7% остварена је уз мерење комплекса различитих лингвистичких маркера (лексичких, синтаксних, семантичких, везаних за садржај текста) [17].

Наредни резултати приказани у табели такође су остварени мерењем комплексног скупа лингвистчких маркера добијених из различитих текстуалних извора које су истраживачи сами прикупили са интернета. Највећи постигнути резултат износи 97.69% за групу од 20 аутора текстова на енглеском језику [22]. У оквиру исте студије, анализа текстова на кинеском за исти број аутора дала је резултат од 88.33%. Високи резултати остварени су и у наредној приказаној студији, која се бави утврђивањем идентитета аутора порука са екстремистичких сајтова и форума - 94.83% за арапски и 97% за енглески језик.

Добијени резултати показују да се употребом *SVM* алгоритама уз мерење комплекса различитих лингвистичких маркера могу постићи значајни резултати на пољу атрибуције ауторства. Наведене студије су показале и велика колебања у резултатима у зависности од низа фактора. Показано је да се применом појединачних лингвистичких маркера (само лексичких или само синтаксних) добијају знатно слабији резултати него када се ти маркери комбинују, односно да језик на коме су текстови написани такође може да утиче на прецизност идентификације [22] [11].

Најбоље остварени резултати у анализираним студијама постигнути су у контролисаним условима, са тачно одређеним кругом кандитата и довољном количином текстуалних узорака за сваког од њих. Поставља се питање ефикасности примене претходно описаних решења у мање контролисаним условима у којима би круг кандидата био или непотпун или знатно већи, односно у којима не би постојао довољан узорак за проучавање стила писања сваког од кандидата. Све то указује да би се резултати идентификације аутора кроз анализу текста требало узети са резервом и да би се требали проверити и потврдити применом других расположивих форензичких метода.

6. ПРАВЦИ ДАЉЕГ РАЗВОЈА

Приступ одабран у овом раду био је да се прикажу пре свега она решења која су постигла најбоље резултате, али која нису опште прихваћена. Даљи рад би из тог разлога усмерити у правцу стандардизације стилометријских решења. Како су чак и она показала велику променљивост резултата у зависности од низа фактора као што су величина скупа података или број аутора, будућа истраживања би требала бити усмерена и ка повећању прецизности модела у мање повољним и мање контролисаним условима [24].

7. ЗАКЉУЧАК

Несумњиво је да ће потреба за оваквом врстом идентификације бити све актуелнија у будућности. Утврђи-

[5] У питању је скуп података који обухвата 50 различитих аутора са по 50 текстова.

вање идентитета аутора кроз анализу текста све више ће се користити у ситуацијама у којима остале форензичке и биометријске методе не буду биле на располагању, односно као допуна за њих. Резултати добијени у досадашњим истраживањима су охрабрујући, али услед осетљивости на велики број фактора показују да стилометрију у погледу поузданости још увек не може да замени ДНК анализу или анализу отисака прстију. Ипак, све веће количине доступног електронског текстуалног материјала, уз све већу практичну потребу за оваквим решењима несумњиво ће представљати подстицај за даљи развој ове области, тако да се може очекивати да ће се она даље усавршавати.

8. ЛИТЕРАТУРА

[1] *Chasky, C.: Who's At The Keyboard? Authorship Attribution in Digital Evidence Investigations, International Journal of Digital Evidence 4(1), 2005, pp.1-13.*

[2] Kašić, Z., Đorđević, P. J.: *Zašto je lingvistika postala forenzička veština. Istraživanja u specijalnoj pedagogiji* (urednik-Dobrivoje Radovanović). Beograd: Fakultet za specijalnu edukaciju i rehabilitaciju, 2009, 469-482. ISBN 978-86-80113-84-5

[3] Черњаков Е.,Б. (1969), Пет векова шпијунаже 1, Београд, Издавачка кућа Народна Књига.

[4] K. Calix, M. Connors, D. Levy, H. Manzar, G. McCabe, and S. Westcott,"Stylometry for E-mail Author Identification and Authentication",Seidenberg School of CSIS, Pace University, New York.

[5] Mendenhall, T. The characteristic curves of composition. *Science*, 214:237249, 1887.

[6] Mendenhall, T. C. 1901. A Mechanical Solution of a Literary Problem. *The Popular Science Monthly* Vol LX: 97–105.

[7] Mosteller, F. and Wallace, D. L. 1964. Inference and Disputed Authorship: The Federalist, Series in Behavioral Science:Quantitative Methods ed. Addison-Wesley, Massachusetts

[8] Nirkhi, Smita. (2014). Stylometric Approach For Author Identification of Online Messages. International Journal of Computer Science and Information Technologies, Vol. 5 (5) , 2014, 6158-6159.

[9] G. UDNY YULE; ON SENTENCE- LENGTH AS A STATISTICAL CHARACTERISTIC OF STYLE IN PROSE: WITH APPLICATION TO TWO CASES OF DISPUTED AUTHORSHIP, *Biometrika*, Volume 30, Issue 3-4, 1 January 1939

[10] George Kingsley Zipf Selected Studies of the Principle of Relative Frequency in Language, by George Kingsley Zipf, Harvard University Press, 1932

[11] Abbasi, A., & Chen, H. (2005). Applying authorship analysis to extremist-group Web forum messages. *IEEE Intelligent Systems*, *20*(5), 67-75. DOI: 10.1109/MIS.2005.81

[12] Chaski, C.E. (2001). Empirical evaluations of language-based author identification techniques.

[13] Frantzeskou, G., Stamatatos, E., Gritzalis, S., & Katsikas, S. (2006). Effective identification of source code authors using byte-level information. In Proceedings of the 28th International Conference on Software Engineering (pp. 893-896).

[14] Shearer C., *The CRISP-DM model: the new blueprint for data mining*, J Data Warehousing (2000); 5:13—22.

[15] Rudman, J. (1998). The state of authorship attribution studies: Some problems and solutions, Computers and the Humanities, 31, 351-365.

[16] Nirkhi, Smita. (2015). Authorship Identification using Generalized Features and Analysis of Computational Method. Transactions on Machine Learning and Artificial Intelligence, Vol. 3 (2) , ISSN 2054-7930.

[17] D'Cruz M., Maximum Likelihood Text Classification Algorithm Using Machine Learning For Authorship Attribution, International Journal of Innovative Research in Computer and Communication Engineering (An ISO 3297: 2007 Certified Organization) Vol. 4, Issue 11, November 2016

[18] Qian, C., He, T., Zhang, R. Deep Learning baseg Authorship Identification, Department of Electrica Engineering, Stanford University, Stanford, CA 94035

[19] Abbasi A, Chen H. Writeprints: A stylometric approach to identity-level identification and similarity detection in cyberspace[J]. ACM Transactions on Information Systems (TOIS), 2008, 26(2): 7.

[20] Li, Jiexun & Zheng, Rong & Chen, Hsiu-chin. (2006). From fingerprint to writeprint. Communications of the ACM 49(4):76-82 · April 2006

[21] *Cortes, Corinna; Vapnik, Vladimir N. (1995). "Support-vector networks". Machine Learning.* ***20*** *(3): 273–297. doi:10.1007/BF00994018*

[22] H. Chen, Z. Huang, J. Li, R. Zheng 2006. A framework for authorship Identification of Online Messages: writing-Style features and classification Techniques. JASIST, pp : 378-393.

[23] A. Anderson, M. Corney, O. DeVel, G. Mohay 2001. Mining E-mail Content for Author Identification Forensics. SIGMOD Record, 30(4), pp : 55-64.

[24] Nirkhi, Smita & Dharaskar, Rajiv. (2013). Comparative study of Authorship Identification Techniques for Cyber Forensics Analysis. International Journal of Advanced Computer Science and Applications. 4. 10.14569/IJACSA.2013.040505.

[25] Dobrović, M., Delić, V., Jakovljević N., Jokić, I.,(2013). *ZAVISNOST TAČNOSTI PREPOZNAVANJA GOVORNIKA OD IZBORA OBELEŽJA*.InfoM 45/2013. udc 378.014.24:005.963(4)

Марко Гогић, студент мастер студија,
Факултет организационих наука Београд
Контакт: m.gogic.fb@gmail.com
Области интересовања: машинско учење, дигитална форензика, базе података