Poređenje tačnosti modela random forest i hibridne neuronske mreže na klasifikaciju teksta u tipove ličnosti po Keirsey modelu

Marko Milenković

Apstrakt - Na (MBTI) Myers-Briggs Personality Type Dataset setu podataka odrađen je preprocessing sa spaCy bibliotekom nakon čega su podaci klasifikovani korišćenjem random forest metode i hibridnog modela neuronskih mreža (CNN+LSTM). Pre treniranja hibridnog modela je odrađen word embedding unapred istreniranim tok2vec alatom koji pruža spaCy. Zbog poređenja je takođe kao referentni model uzet hibridni model koji nije koristio spaCy word embedding nego tensorflow word embedding layer. Random forest model je imao najveći prosečan f1-score 70% sa 10000 stabla, a najmanji prosečan f1-score 57% sa 100 stabla. Hibridni model sa spaCy word embeddingom je imao najveći prosečan f1-score 65%, a referentni model je imao maksimalni prosečan f1-score 75%. U radu se pokazalo da je hibridni model precizniji kao i da unapred istreniran spaCy word embedding nije adekvatan za primenu na objave koje sadrže puno internet žargona.

I. Uvod

Temperament, kao i ostale klasifikacije ličnosti, se obično određuju ispunjavanjem formulara. Ovo je nepraktično jer zahteva puno vremena i ne može se očekivati od svakoga da izdvoji ponekad i sat vremena za to. Takođe, takav način podrazumeva samoprocenjivanje, što nije uvek precizno. Poslednjih godina je procena ličnosti postala popularna i zato je interesantna tema za mnogo istraživača. Motivacija iza ovog rada je da se pronađe pouzdan način aproksimacije vrste temperamenta koju ima neki pojedinac samo analizom njegovih objava, odnosno, načina pisanja. Da bi se to postiglo mogu se primeniti NLP (*natural language processing*) tehnike [1] i machine learning algoritmi. Za klasifikaciju temperamenta je korišćena podela Davida Keirseya [2] zbog njene praktičnosti i kompatibilnosti sa MBTI klasifikacijom ličnosti [3].

RF (*random forest*) model [4] je supervizovani algoritam za mašinsko učenje [5] koji se koristi za klasifikaciju i regresiju. Prednost je što se jako lako implementira i lako menja.

M. Milenković je učenik 4. IT odeljenja Gimnazije „Svetozar Marković”, Petefi Šandora 1, 24000, Subotica, Srbija / polaznik seminara računarstva u Istraživačkoj stanici Petnica, Selo Petnica, 14104 Valjevo, Srbija, E-mail: mmarkomile@gmail.com

Ali, zbog svoje jednostavnosti se ne može očekivati da daje najpreciznije rezultate kada su u pitanju komplikovani zadaci. RF modeli se koriste u raznim situacijama - predviđanje berze, dijagnostika u zdravstvu, psihologija, sistemima za preporuku i dr. Neuronske mreže [6] se takođe mogu koristiti u jako puno različitih situacija, pri čemu se koriste različite vrste neuronskih mreža. Tako su i počeli nastajati hibridni modeli u kojima se različite vrste neuronskih mreža sekvencijalno spajaju kako bi se poboljšala njihova preciznost. Hibridni model koji se u ovom radu koristio je spoj CNN (*convolutional neural network* - konvoluciona neuronska mreža) [7] i LSTM (*long short term memory*) neuronske mreže [8]. Razlog zašto je to dobra kombinacija je jer CNN dobro izdvaja bitne delove rečenica, a LSTM efektivno povezuje te delove i pronalazi kontekst.

U prošlim radovima su predstavljeni razni modeli koji uglavnom koriste NLP metode namenjene za istraživanje i naučne radove ili metode koje su prevaziđene [9]-[12],[17]. U ovom istraživanju je korišćena spaCy biblioteka [13] koja je namenjena da bude laka za implementiranje i dostupna svim programerima. Nakon preprocesiranja teksta iz dataseta, tekst se lematizovao alatom iz spaCy biblioteke i onda se od njega pravio *bag of words* u slučaju RF modela. (lematizacija i *bag of words* su objašnjeni u narednom delu rada) U slučaju hibridnog modela se koristio unapred istreniran tok2vec alat iz spaCy biblioteke koji pretvara reči u jedinstvene 300-dimenzionalne vektore koji ih predstavljaju, nakon čega se na njima trenirao i evaluirao model. Cilj je da se proveri da li spaCy biblioteka pruža dovoljan nivo preciznosti i primene, kao i da se poredi RF model sa hibridnim modelom baziranim na neuronskim mrežama.

II. Metode za klasifikaciju

A. Dataset

Korišćen je (MBTI) Myers-Briggs Personality Type Dataset [14] koji sadrži preko 8600 redova podataka. Svaki red sadrži, u jednoj koloni tip ličnosti (jedan od 16 MBTI tipova ličnosti), a u drugoj koloni poslednjih 50 objava koju je ta osoba objavila, razdvojenih sa „|||” (3 uspravne crte). Podaci su uzeti sa PersonalityCafe foruma.

B. Preprocesiranje

Pošto korisnici društvenih mreža i foruma često koriste neformalni vokabular i različite simbole koji mogu biti korisni za klasifikaciju ličnosti, ali stvaraju velike poteškoće modelima mašinskog učenja i zauzimaju mnogo memorije potrebno je očistiti dataset. Kako bi se pročistio dataset niz operacija je izvršeno na njemu pre nego što je korišćen za treniranje algoritama. Operacije koje se izvršavaju su uklanjanje uspravnih crta, zamena linkova za njihov domen, uklanjanje viška razmaka, uklanjanje objava koje sadrže samo brojeve ili linkove, uklanjanje specijalnih i akcentovanih karaktera, uklanjanje stopwords-a, uklanjanje brojeva, pretvaranje u mala slova, sklanjanje ponavljanih karaktera, uklanjanje reči dužine 1 karaktera i uklanjanje najčešćih i najređih reči. Takođe, izdvojeni su podaci koji bi mogli biti korisni RF modelu: broj reči napisanih velikim slovima (u celini), broj stopwords-a, broj linkova, broj reči, prosečna dužina reči.

Zato što je korišćen dataset namenjen za MBTI klasifikaciju, bilo je potrebno mapirati MBTI klase u Keirsey model. „ISTJ" „ISFJ" „ESFJ" „ESTJ" su mapirani u 0 (Guardian), „ISFP" „ISTP" „ESFP" „ESTP" su mapirani u 1 (Artisan), „INFJ" „INFP" „ENFP" „ENFJ" su mapirani u 2 (Idealist), „INTJ" „INTP" „ENTP" „ENTJ" su mapirani u 3 (Rationalist).

C. Primena spaCy biblioteke

spaCy je natural language processing biblioteka za Python. Namenjena je za programere i napravljena je tako da je jednostavna za korišćenje. Korišćena je za lematizaciju [15] reči nakon prvog dela preprocesiranja. Lematizacija reči je proces u kojem se reči pretvaraju u svoj koren (lemu). To je korisno jer drastično smanjuje količinu podataka i gube se nepotrebne razlike između reči koje u osnovi nose isto značenje, ali su različitog oblika. Korišćenjem alata koji pruža biblioteka preprocesirane reči se zamenjuju njihovim lemama.

D. Random forest model

Kako bi se uzeo ujednačen uzorak sve 4 klase, nasumično se odabralo 450 objava od svake klase koje će se koristiti za treniranje modela. Da bi objave koje se nalaze u datasetu mogle da se koriste za učenje RF modela, prvo se pravi *bag of words* [16] svih objava. Ovako se pravi vektor celih brojeva dužine broja jedinstvenih reči u svim objavama koji predstavljaju da li i koliko puta se neka reč nalazi u objavi. *Bag of words* vektor se spaja sa ostalim izdvojenim podacima dobijenim tokom preprocesiranja i dobijeni dataset se koristio za treniranje i evaluaciju. Model je treniran sa 100, 200, 500, 1000 i 10000 stabla, a za treniranje je korišćeno 80% odabranih objava. Ograničenja dubine nisu postavljena.

E. Hibridni deep learning model

Hibridni model koji je evaluiran je hibridni model pokazan u istraživanju A Hybrid Deep Learning Technique for Personality Trait Classification From Text [17]. Prvi sloj modela u pomenutom istraživanju je *embedding layer*. Njegova uloga je da pretvori reči u vektore koji mogu da se koriste u narednim delovima modela. On se iterativno poboljšava kao i neuronske mreže da bi bolje predstavio reči vektorima. *Embedding layer* tog modela je zamenjen unapred istreniranim spaCy tok2vec alatom. Tok2vec od reči u stringu (tokena) pravi 300-dimenzionalne vektore koji predstavljaju tu reč u vektorskom prostoru. Ovom metodom svaka reč iz spaCy vokabulara može da se predstavi jedinstvenim vektorom. Da bi se dobijena lista vektora mogla koristiti za treniranje odrađen je padding (na kraju svake liste dodat je niz nula vektora) kako bi sve liste bile iste dužine. Odrađen je i one hot encoding labela u 4 dimenzionalni vektor.

Takođe, za razliku od spomenutog modela, korišćen je Adam [18] optimizer jer je pokazao malo bolje rezultate od Adamax [18]. LSTM mreža je trenirana sa 120 jedinica.

Kao referentni model uveden je i model koji ne koristi spaCy biblioteku, nego sadrži embedding sloj u sebi (ovaj model je skoro identičan kao u ranije spomenutom istraživanju).

F. Evaluacija modela

Za evaluaciju tačnosti modela gledao se *f1-score*, *preciznost* i *recall* za svaku od klasa. Ovo je bilo potrebno zbog toga što nisu u svim slučajevima korišćeni uravnoteženi datasetovi i zbog toga što sve te metrike zajedno daju kompletniju sliku performansi modela.

Tabela 1. Random forest model (uzorak od 450)

broj stabla	f1-score (%)				recall (%)				preciznost (%)			
	0	1	2	3	0	1	2	3	0	1	2	3
100	56	59	53	6o	53	61	53	61	60	57	53	59
200	59	61	58	64	54	66	56	68	65	58	61	60
500	67	65	61	67	61	68	59	72	75	62	64	62
1000	69	68	66	65	60	73	67	69	82	64	66	62
2000	72	68	69	69	63	74	69	71	83	63	69	67
10000	71	68	69	68	63	73	69	70	80	63	69	66

III. REZULTATI I DISKUSIJA

U rezultatima RF modela (tabela 1) se vidi da se povećavanjem broja stabla, povećava i *f1-score*. *F1-score*

od 70% se računa kao prosek sa 10000 stabla i predviđen je kao gornja granica mogućnosti RF modela.

Hibridni model je ispitan u 2 okolnosti. U prvom slučaju je nasumično odabrano 450 uzoraka od svake klase i to je korišćeno kao set za treniranje i validaciju. U drugom slučaju korišćen je ceo dataset. Ovo služi da poredimo koliko je bitna ravnomerna raspodela klasa. Po rezultatima hibridnog modela (tabela 2 i tabela 3) možemo videti da su modeli imali bolji *f1-score* kada se koristi ceo dataset. Takođe se primećuje značajno veća preciznost modela koji ne koristi spaCy biblioteku (tabela 3) u oba slučaja. Za sve slučajeve neuronska mreža dostiže svoju maksimalnu preciznost oko 10 epoha, nakon čega počinje da opada ili stagnira zbog overfitovanja. Ovo se možda može povezati sa ograničenjima dataseta. Maksimalni prosečni *f1-score* RF modela je 70% sa 10000 stabla pri korišćenju uzoraka od 450 redova. Maksimalni prosečni *f1-score* neuronske mreže je iznosio 75% u modelu bez spaCy biblioteke i 65% u modelu sa spaCy bibliotekom, u oba slučaja korišćenjem celog dataseta.

Bitno je napomenuti da se govori o *weighted average* načinu računanja proseka. To podrazumeva proporcije labela u datasetu. U slučaju *macro average* načina, koji ne uvažava proporcije labela rezultati su drugačiji. Ovo se da primetiti u modelu sa spaCy bibliotekom koji je koristio ceo dataset (tabela 2).

Tabela 2. Hibridni model + spaCy

dataset	f1-score (%)				recall (%)				preciznost (%)			
	0	1	2	3	0	1	2	3	0	1	2	3
uzorci od 450	42	38	51	47	49	31	53	46	36	48	49	49
ceo dataset	0	1	72	66	0	1	81	68	0	33	65	64

Ne gledajući na proporcije, ovaj model izgleda kao da ima jako loš prosek (35%) jer je zbog neravnomerno raspoređenih klasa u datasetu skoro sasvim zanemario klase 0 i 1.

IV. Zaključak

U rezultatima se može videti da je najbolji metod hibridni model bez korišćenja spaCy biblioteke. Ovo je najverovatnije slučaj zbog toga što internet žargon sadrži jako širok spektar reči koje se ne nalaze u spaCy vokabularu, pa ne može napraviti odgovarajuće vektore.

Takođe je bitno napomenuti da je dataset prilično ograničen. Sadrži samo oko 8600 osoba što je veoma malo za potrebe neuronskih mreža. Veliki broj objava ne sadrži značajne podatke što još više smanjuje preciznost modela. Veoma su neujednačeni uzorci ličnosti jer se klasa 0 pojavljuje samo oko 450 puta, a klasa 3 čak oko 4500 puta, zbog ovoga su modeli naklonjeni ka češćim klasama.

To je i bio razlog za biranje uzoraka od 450 u testiranju. Pokazalo se da je ipak bolje koristiti ceo dataset, bez obzira na to što su klase neravnomerno raspoređene, ali treba obratiti pažnju da model ne prestane razmatrati one klase koje se retko pojavljuju.

U daljem istraživanju bi se mogao sličan ogled odraditi na malo boljem datasetu. Iako su očigledna ograničenja dataseta, unapred istreniran spaCy tok2vec alat je pokazao da nije adekvatan za primenu na objave na internetu koje sadrže veliku količinu internet žargona. spaCy je pored toga bio sasvim adekvatan kao alat za preprocesiranje teksta, odnosno pretvaranja reči u njihove leme.

Tabela 3. Hibridni model bez spaCy

dataset	f1-score (%)				recall (%)				preciznost (%)			
	0	1	2	3	0	1	2	3	0	1	2	3
uzorci od 450	66	55	47	64	69	66	34	64	63	47	72	63
ceo dataset	50	50	82	76	58	42	88	70	43	63	76	83

Reference

[1] Chowdhary, K., 2020. Natural language processing. *Fundamentals of artificial intelligence*, pp.603-649.

[2] Keirsey, D. & Bates, M.M., 1978. *The sixteen types*, Prometheus Nemesis Book Co.

[3] Myers, I.B., 1962. The Myers-Briggs Type Indicator: Manual (1962).

[4] Breiman, L., 2001. Random forests. *Machine learning*, 45(1), pp.5-32.

[5] Kotsiantis, S.B., Zaharakis, I. and Pintelas, P., 2007. Supervised machine learning: A review of classification techniques. *Emerging artificial intelligence applications in computer engineering*, 160(1), pp.3-24.

[6] Bishop, C.M., 1994. Neural networks and their applications. *Review of scientific instruments*, 65(6), pp.1803-1832.

[7] Albawi, S., Mohammed, T.A. and Al-Zawi, S., 2017, August. *Understanding of a convolutional neural network*. In *2017 international conference on engineering and technology (ICET)* (pp. 1-6). Ieee.

[8] Sherstinsky, A., 2020. Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network. *Physica D: Nonlinear Phenomena, 404*, p.132306.

[9] Arnoux, P.H., Xu, A., Boyette, N., Mahmud, J., Akkiraju, R. and Sinha, V., 2017, May. 25 tweets to know you: A new model to predict personality with social media. In Eleventh international AAAI conference on web and social media.

[10] Keh, S.S. and Cheng, I., 2019. Myers-Briggs personality classification and personality-specific language generation using pre-trained language models. arXiv preprint arXiv:1907.06333.

[11] Lima, A.C.E. and de Castro, L.N., 2019. TECLA: A temperament and psychological type prediction framework from Twitter data. Plos one, 14(3), p.e0212844.

[12] Wang, X., Sui, Y., Zheng, K., Shi, Y. and Cao, S., 2021. Personality classification of social users based on feature fusion. Sensors, 21(20), p.6758.

[13] Honnibal, M., Montani, I., Van Landeghem, S. and Boyd, A., 2020. spaCy: Industrial-strength natural language processing in python.

[14] J, M., 2017. (MBTI) Myers-Briggs personality type dataset. *Kaggle*. Available at: https://www.kaggle.com/datasets/datasnaek/mbti-type [Accessed September 25, 2022].

[15] Balakrishnan, V. and Lloyd-Yemoh, E., 2014. Stemming and lemmatization: A comparison of retrieval performances.

[16] Zhang, Y., Jin, R. and Zhou, Z.H., 2010. Understanding bag-of-words model: a statistical framework. *International journal of machine learning and cybernetics, 1*(1), pp.43-52.

[17] Ahmad, H., Asghar, M.U., Asghar, M.Z., Khan, A. and Mosavi, A.H., 2021. A hybrid deep learning technique for personality trait classification from text. *IEEE Access, 9*, pp.146214-146232.

[18] Kingma, D.P. and Ba, J., 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.