Fake news challenge - Detekcija lažnih vijesti

Fran Mišić
University of Zagreb, Faculty of Science, Department of Mathematics

Andrej Slapničar
University of Zagreb, Faculty of Science, Department of Mathematics

Iva Sokolaj
University of Zagreb, Faculty of Science, Department of Mathematics

Roko Torbarina
University of Zagreb, Faculty of Science, Department of Mathematics

Fake News Challenge was a competition with task to classify stances between article headlines and article bodies. In this paper, we analyze predictive models of top three teams and create our own models based on decision trees and BERT entailment.

Keywords—fake news, term frequency-inverse document frequency, Word2Vec, decision trees, multilayer perceptron, Bidirectional Encoder Representations from Transformers

I. UVOD

Fake news challenge [1] trebao je biti niz natjecanja s ciljem istraživanja upotrebe umjetne inteligencije u problemu detekcije lažnih vijesti. Razvojem umjetne inteligencije moći će se sve više i više automatizirati dijelovi procedure za provjeru lažnih vijesti koju trenutno ljudi ručno rade. Dosad se održalo samo jedno natjecanje: FNC-1. FNC-1 održan je 2016. godine te se fokusirao na zadatak detekcije stajališta. Pod detekcijom stajališta misli se na istraživanje odnosa između dva teksta. U FNC-1 istražuje se mogućnost procjene odnosa teksta iz novinskog članka s obzirom na naslov. Članak se može slagati, ne slagati, raspravljati ili biti nepovezan s naslovom.

II. PODATCI I BODOVANJE

Na natjecanju je sudjelovalo 50 timova. Timovima je bio predstavljen skup podataka (*dataset*) za treniranje od 49972 instanci koje se sastoje od naslova vijesti, sadržaja vijesti te odnosa naslova i sadržaja. Odnos može biti jedan od: *agree*, *disagree*, *discuss* i *unrelated*. Također je bio predstavljen i skup podataka za testiranje koji se sastoji 25413 instanci naslova i sadržaja vijesti.

Naslov vijesti	Explosion reported near the Nicaraguan capital attributed to meteorite impact, but experts, including NASA, cast doubt on claims
Sadržaj vijesti	News has been circulating about a potential meteorite strike near Managua, Nicaragua late Saturday night, just 13 hours or so before the close flyby of 20-m asteroid 2014 RC, leading some to suggest that the two events are related. ⋮ While this particular event is looking more and more like a false alarm with time, it should be noted that fireballs blaze through our skies every day, as tons of material is swept up by Earth as the planet orbits the Sun. Many of these are missed because they occur during the day, or over regions of the planet that aren't heavily populated.
Odnos naslova i sadržaja	agree

Fig. 1. Primjer jedne instance iz skupa podataka.

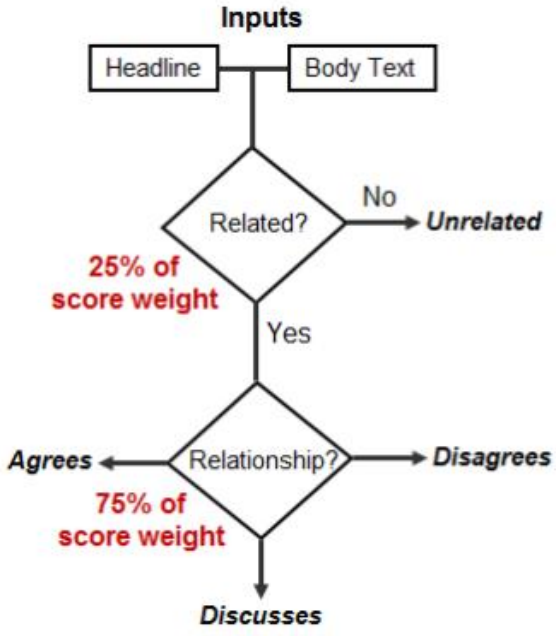

Fig. 2. Bodovanje uspješnosti modela.

[1] http://www.fakenewschallenge.org/

Konačni rezultati bodovali su se na skupu podataka od 25413 parova naslov-sadržaj prema idućim smjernicama:

- +0.25 bodova ako je točno određena povezanost (*unrelated* /*related*)
- +0.75 bodova ako je određena točna povezanost (*agree* / *disagree* / *discuss*).

Od 25413 instance, njih 7064 bilo je povezano, a 18349 nepovezano. Dakle, maksimalan broj bodova bio je:

$$7064 \cdot 1 + 18349 \cdot 0.25 = 11651.25.$$

III. METODE I REZULTATI NAJBOLJIH TIMOVA NA NATJECANJU

A. *Značajke (features)*

Sva tri tima na natjecanju su koristili slične značajke. U nastavku su navedena [2] objašnjenja značajki koje se pojavljuju u njihovim modelima:

- Kosinusna sličnost (*cosine similarity*) između vektora naslova i vektora dokumenta koji su bazirani na produktu frekvencije riječi i inverzne frekvencije dokumenata (*term frequency-inverse document frequency* – TF-IDF).

  Kosinusna sličnost pokazuje koliko su dva vektora slična, odnosno koliki je kut između ta dva vektora.

$$\text{cosine similarity} = S_C(A,B) := \cos(\theta) = \frac{\mathbf{A}\cdot\mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$$

Fig. 3. Formula kosinusne sličnosti vektora *A* i *B*.

  Produkt frekvencije riječi i inverzne frekvencije dokumenata (TF-IDF) je mjera koja pokazuje koliko je neka riječ bitna za tekst određenog dokumenta unutar kolekcije dokumenata. TF–IDF se računa kao umnožak frekvencije riječi u dokumentu i inverzne frekvencije riječi u svim dokumentima.

$$\mathrm{tf}(t,d) = \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}}$$

$$\mathrm{idf}(t,D) = \log \frac{N}{|\{d \in D : t \in d\}|}$$

$$\mathrm{tfidf}(t,d,D) = \mathrm{tf}(t,d) \cdot \mathrm{idf}(t,D)$$

Fig. 4. Formula *tf(t, d)* frekvencije riječi *t* u dokumentu *d*, formula *idf(t, D)* inverzne frekvencije riječi *t* u skupu dokumenata *D* te formula *tfidf(t, d, D)* produkta frekvencije riječi *t* u dokumentu *d* i inverzne frekvencije riječi *t* u skupu dokumenata *D*.

- Latentna Dirichletova alokacija.

  LDA je statistička tehnika klasifikacije objekata u međusobno isključive grupe bazirane na mjerenim svojstvima objekata. Pri primjeni pazi se na dvije glavne točke: koja svojstva objekta će odrediti pripadnost pojedine grupe i koji model ili pravilo najbolje razlučuje pojedine grupe.

- Latentno semantičko indeksiranje.

  LSI metoda je indeksiranja i pronalaženja koja koristi rastavljanje singularne vrijednosti (SVD) za prepoznavanje uzoraka u odnosima između pojmova i koncepata sadržanih u nestrukturiranoj zbirci teksta. Temelji se na načelu da riječi koje se koriste u istom kontekstu obično imaju slična značenja.

- Frekvencije unigrama, bigrama i trigrama.

  n-gram je uzastopni podniz od n riječi iz zadanog uzorka teksta. Primjerice u rečenici „Jučer sam išao u kino.“:

  - unigrami (1-gram) su „Jučer“, „sam“ „išao“, „u“ i „kino“;
  - bigrami (2-gram) su „Jučer sam“, „sam išao“ „išao u“ i „u kino“;
  - trigrami (3-gram) su „Jučer sam išao“, „sam išao u“ i „išao u kino“.

- Singularna dekompozicija (*singular value decomposition*) TF–IDF vektora

  Singularna dekompozicija (SVD) je faktorizacija matrice M u obliku $M = U\Sigma V^*$, gdje je U kvadratna unitarna matrica, Σ

[2]

dijagonalna matrica s nenegativnim vrijednostima i V je kvadratna unitarna matrica.

- Word2Vec vektori riječi (*word embeddings*)

  Word2vec[3] je NLP metoda napravljena 2013. godine. Kreirao ju je tim znanstvenika u Googleu na čelu kojeg je bio Tomas Mikolov. Word2Vec je model koja koristi neuronsku mrežu kako bi naučio asocijacije među riječima iz velikog broja tekstova. Korišten je predtrenirani Word2Vec model koji svaku riječ reprezentira preko 300-dimenzionalnog vektora tako da su vektori semantički sličnih riječi i sami slični, pa koristeći neku mjeru sličnosti vektora kao što je kosinusna sličnost može se utvrditi jesu li riječi slične.

- Sentiment.

  U analizi sentimenta cilj je utvrditi polaritet mišljenja zasebno u naslovu i sadržaju vijesti. Negativni polaritet znači da tekst iskazuje negativno mišljenje o nekoj temi, a pozitivni polaritet ima tekst koji iskazuje pozitivno mišljenje. Sami polaritet ne govori ništa o tome govore li naslov i sadržaj o istoj temi, ali zato se koristi zajedno s drugim značajkama.

B. Model tima UCL Machine Reading

Treće mjesto na FNC osvojio je tim UCL *Machine Reading*. Korišten je klasifikator zasnovan na višeslojnom perceptronu (*multilayer perceptron* - MLP) s jednim skrivenim slojem od 100 jedinica i sa *softmax*-om na izlazu konačnog linearnog sloja.

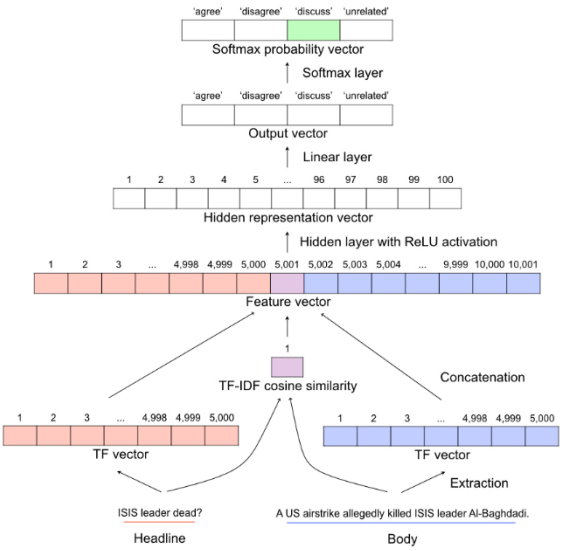

Fig. 5. Shematski prikaz UCL modela.

Za nelinearnost skrivenog sloja iskorištena je ReLU aktivacijska funkcija (rectified linear unit).

$$f(x) = x^+ = \max(0, x)$$

Fig. 6. Formula RELU aktivacijske funkcije.

Svakoj oznaci *agree*, *disagree*, *discuss* i *unrelated* pridružen je broj bodova, te je kao konačni rezultat izbačena oznaka s najvećim brojem bodova. Kao značajke (*feature*), uzeti su vektori frekvencija 5000 najčešćih riječi u skupu podataka. Također, korištena je kosinusna sličnost (*cosine similarity*) između vektora naslova i vektora dokumenta koji su bazirani na produktu frekvencije riječi i inverzne frekvencije dokumenata (*term frequency-inverse document frequency* – TF-IDF).

Cilj treniranja bio je minimalizirati unakrsnu entropiju između *softmax* vjerojatnosti i stvarnih oznaka. Treniralo se u malim serijama s povratnim širenjem (*backpropagation*) pomoću Adam-optimizatora i gradijentnog rezanja prema globalnom *norm clip* omjeru.

$$\sigma : \mathbb{R}^K \to (0,1)^K$$

$$\sigma(\mathbf{z})_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}} \quad \text{for } i = 1, \ldots, K \text{ and } \mathbf{z} = (z_1, \ldots, z_K) \in \mathbb{R}^K.$$

Fig. 7. Formula *softmax* funkcije vektora *z*.

[3] https://code.google.com/archive/p/word2vec/

C. *Model tima Athene (UKP Lab)*

Tim Athene je koristio klasifikaciju zasnovanu na višeslojnom perceptronu (MLP) nadograđujući rad Davis i Proctora (2017).

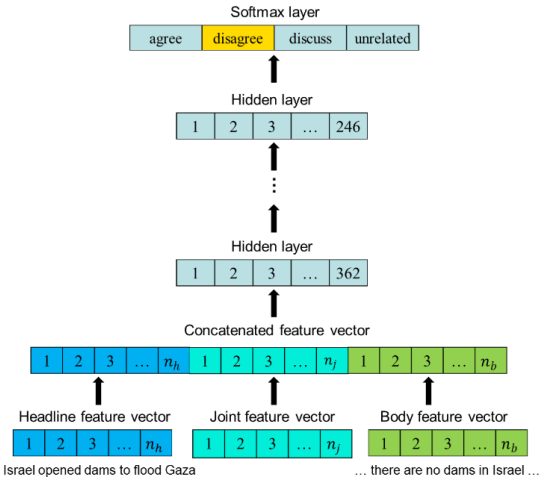

Fig. 8. Shematski prikaz Athene modela.

Njihov model sastoji se od 6 skrivenih slojeva i *softmax* sloja. Koriste se mnoge značajke: unigrami, kosinusna sličnost (*cosine similarity)*, latentna Dirichletova alokacija (*latent Dirichlet allocation* – LDA), latentno semantičko indeksiranje (*latent semantic indexing)*. Ovisno o tipu značajki, stvara se: ili zajednički vektor značajki (*feature vector*) za naslov i sadržaj, ili posebni vektor značajki za naslov i vektor značajki za sadržaj, te se onda ta dva vektora konkateniraju.

D. *SOLAT in the SWEN (Talos)*

Prvo mjesto ostvario je tim Talos, njihov model rezultate donosi na temelju aritmetičke sredine predikcija dobivenih metodom stabla odlučivanja s pojačanim gradijentom (*gradient-boosted decision trees* - GBDT) i dubokim konvolucijskim neuronskim mrežama (*deep convolutional neural network - deep* CNN).

Model stabla odlučivanja s pojačanim gradijentom (GBDT) uzima nekoliko tekstualno-baziranih značajki iz naslova i sadržaja vijesti: frekvencije unigrama, bigrama i trigrama, produkt frekvencije riječi i inverzne frekvencije dokumenata (TF–IDF), singularna dekompozicija (*singular value decomposition* - SVD) TF–IDF vektora, Word2Vec vektore riječi (*word embeddings*) i sentiment. Korištena je XGBoost[4] implementacija ručno podešena za navedene značajke vijesti.

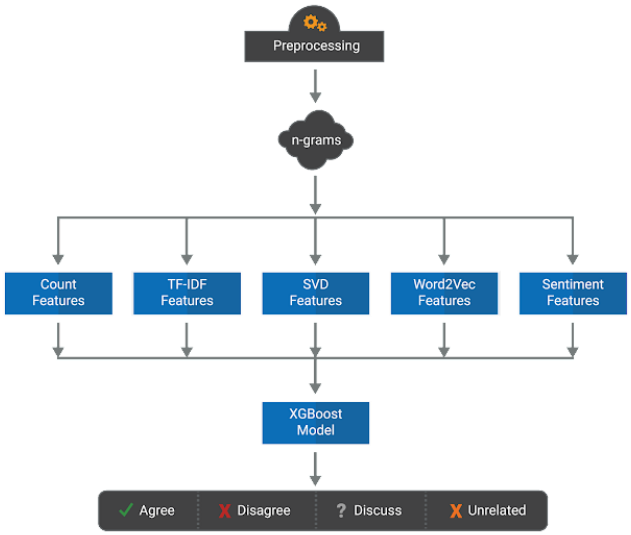

Fig. 9. Shematski prikaz Athene GBDT modela.

Model s dubokim konvolucijskim neuronskim mrežama (*deep* CNN) koristi jednodimenzionalnu konvolucijsku mrežu (1D-CNN) na naslovu i sadržaju vijesti, koji se reprezentiraju preko Googleovog predtreniranog Word2Vec modela. Zatim se na izlazu koji daje 1D-CNN trenira višeslojni perceptron (MLP) koji kao rezultat daje odnos naslova i sadržaja vijesti (*agree*, *disagree*, *discuss* ili *unrelated*).

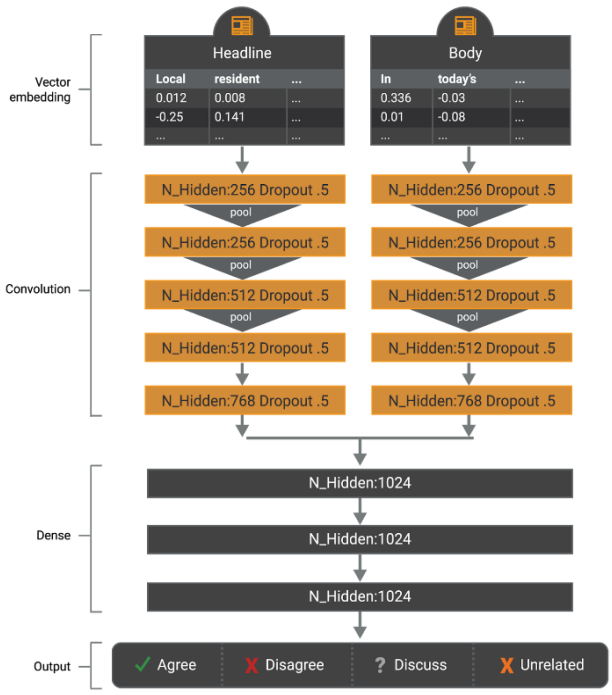

Fig. 10. Shematski prikaz Athene *deep* CNN modela.

[4] https://xgboost.ai/

E. *Uspjeh timova*

Prikaz uspješnosti modela timova UCL, Athene i Talos na natjecateljskom skupu podataka za testiranje dan je u tablicama prikazanim na slikama Fig. 11, Fig. 12 i Fig. 13. Prvu nagradu (USD 1000) na natjecanju osvojio je tim SOLAT in the SWEN (Talos) s ukupnih 9556.50 bodova, što je 82.02 % maksimalnog broja bodova.

	Agree	Disagree	Discuss	Unrelated	Accuracy (%)
Agree	838	12	939	114	44.04
Disagree	179	46	356	116	6.60
Discuss	523	46	3633	262	81.38
Unrelated	53	3	330	17963	97.90

Fig. 11. Rezultati tima UCL na test *dataset*-u iz natjecanja.

	Agree	Disagree	Discuss	Unrelated	Accuracy (%)
Agree	851	69	826	157	44.72
Disagree	241	66	241	149	9.47
Discuss	466	37	3611	350	80.89
Unrelated	19	4	115	18211	99.25

Fig. 12. Rezultati tima Athene na test *dataset*-u iz natjecanja.

	Agree	Disagree	Discuss	Unrelated	Accuracy (%)
Agree	1114	17	588	184	58.54
Disagree	275	13	294	115	1.87
Discuss	823	6	3401	234	76.19
Unrelated	35	0	203	18111	98.70

Fig. 13. Rezultati tima Talos na test *dataset*-u iz natjecanja.

Tablica prikazana na slici Fig. 14 sadrži ukupni i relativni broj bodova za prva tri tima na natjecanju.

Rank	Team name	Score	Relative Score
1	SOLAT in the SWEN	9556.50	82.02
2	Athene (UKP Lab)	9550.75	81.97
3	UCL Machine Reading	9521.50	81.72

Fig. 14. Konačna rang lista prva tri tima na natjecanju.

F. *Greške u predikciji*

Podjela krivo klasificiranih primjera u skupu podataka za testiranje po timovima je prikazana u Vennovom dijagramu na slici Fig. 15.

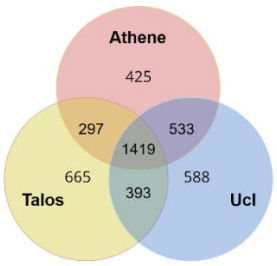

Fig. 15. Vennov dijagram krivo klasificiranih primjera po timovima.

Modeli rade slične greške, odnosno znatan broj primjera sva tri modela klasificiraju pogrešno. Najčešći razlozi pogrešnih klasifikacija su:

- Podudaranje riječi između naslova i članka može nepovezane parove klasificirati kao povezane.

TABLE I. PRIMJER GREŠKE PRVOG TIPA

Naslov	Saudi Arabia's national airline to introduce gender segregation after a string of complaints from male passengers			
Sadržaj	Saudi women with attractive eyes may be forced to cover them up...			
Odnos	Talos	Athene	UCL	Stvarni
	discuss	*discuss*	*discuss*	*unrelated*

- Korištenje dva sinonima istog pojma u naslovu i sadržaju može povezane parove klasificirati kao nepovezane.

TABLE II. PRIMJER GREŠKE DRUGOG TIPA

Naslov	3-Boobed Woman a Fake			
Sadržaj	She made headlines around the world when she revealed she paid thousands of dollars to get a third breast...			
Odnos	Talos	Athene	UCL	Stvarni
	unrelated	*unrelated*	*unrelated*	*agree*

- Pojavljivanje određenih riječi kao što su “allegedly“, “according to“ i “said“, u sadržaju može rezultirati pogrešnom *discuss* klasifikacijom.

TABLE III. PRIMJER GREŠKE TREĆEG TIPA

Naslov	'How's it going?': Teenager wakes up during brain surgery and asks doctors for progress report			
Sadržaj	Halfway through brain surgery aimed to remove a cancerous growth, a teenager allegedly woke up and asked the doctors...			
Odnos	Talos	Athene	UCL	Stvarni
	discuss	*discuss*	*discuss*	*agree*

Sva tri tima najslabiji uspjeh imali su s *disagree* klasifikacijom. U skupu podataka za testiranje je samo jedna instanca koju su sva 3 modela točno klasificirali kao *disagree*. Lošoj klasifikaciji doprinijeli su mali broj instanci s odnosom *disagree* u skupu podataka i korištene značajke nisu pogodne za određivanje razlika između *agree*, *disagree* i *discuss*.

IV. KORIŠTENE METODE I REZULTATI

A. Klasifikacija koristeći stabla odlučivanja

U ovom jednostavnijem module korištena je metoda stabla odlučivanja za klasifikaciju odnosa naslova i sadržaja vijesti. Metoda Stabla odlučivanja daje klasifikacijski ili regresijski prediktivni model kojeg predstavlja stablo u kojem čvorovi sadrže testove koji se ispituju na značajkama, grane predstavljaju ishod tih testova te listovi sadrže klasifikacijske oznake. Za značajke su uzeti: kosinusna sličnost aritmetičke sredine Word2Vec vektora riječi u naslovu i sadržaju, kosinusna sličnost vektora produkta frekvencije riječi i inverzne frekvencije dokumenata (TF-IDF vektora riječi) u naslovi i sadržaju vijest, zbroj frekvencija imena iz naslova u sadržaju vijesti. Korištena je implementacija klasifikatora sa stablima odlučivanja iz Python biblioteke Scikit-learn[5].

Prikaz uspješnosti modela s Word2Vec značajkama, modela s TF-IDF značajkama te zajednički model s Word2Vec i TF-IDF značajkama na natjecateljskom skupu podataka za testiranje dan je u tablicama prikazanim na slikama Fig. 16, Fig. 17 i Fig. 18. Word2Vec model postiže uspješnost od 65.95 %, TF-IDF model postiže uspješnost 73.03 %, a zajednički model postiže uspješnost 74.27 %.

	Agree	Disagree	Discuss	Unrelated	Accuracy (%)
Agree	264	47	558	307	22.45
Disagree	41	22	109	72	9.02
Discuss	532	115	1416	641	52.37
Unrelated	276	82	673	9837	90.51

Fig. 16. Rezultati Word2Vec modela.

	Agree	Disagree	Discuss	Unrelated	Accuracy (%)
Agree	283	70	703	120	24.06
Disagree	60	16	152	16	6.56
Discuss	712	164	1600	228	59.17
Unrelated	85	24	231	10528	96.87

Fig. 17. Rezultati TF-IDF modela.

	Agree	Disagree	Discuss	Unrelated	Accuracy (%)
Agree	326	54	702	94	27.72
Disagree	60	31	137	16	12.70
Discuss	676	180	1639	209	60.61
Unrelated	84	29	223	10532	96.91

Fig. 18. Rezultati zajedničkog Word2Vec - TF-IDF modela.

Kako Word2Vec model daje slabije rezultate od jednostavnog TF-IDF modela pokušali smo popraviti uspješnost Word2Vec model dodavanjem značajke zbroj frekvencija imena iz naslova u sadržaju vijesti. Promatrane su riječi iz naslova koje Word2Vec ne prepoznaje kao standardne engleske riječi, pretpostavljamo da će to biti imena i nazivi, te je izračunat zbroj frekvencija tih riječi u sadržaju vijesti. Uspješnosti modela s frekvencijama imena i zajedničkog modela s frekvencijama imena Word2Vec značajkama na natjecateljskom skupu podataka za testiranje prikazane su u tablicama na slikama Fig. 19, Fig. 20 i Fig. 21.

	Agree	Disagree	Discuss	Unrelated	Accuracy (%)
Agree	171	23	217	765	14.54
Disagree	21	14	49	160	5.74
Discuss	199	46	649	1810	24.00
Unrelated	28	8	28	10804	99.41

Fig. 19. Rezultati NamesFrequency modela.

	Agree	Disagree	Discuss	Unrelated	Accuracy (%)
Agree	294	59	603	220	25.00
Disagree	54	24	111	55	9.84
Discuss	589	126	1476	513	54.59
Unrelated	202	73	546	10047	92.45

Fig. 20. Rezultati zajedničkog Word2Vec - NamesFrequency modela.

	Agree	Disagree	Discuss	Unrelated	Accuracy (%)
Agree	357	63	662	94	30.36
Disagree	61	27	136	20	11.07
Discuss	697	137	1663	207	61.50
Unrelated	90	30	219	10529	96.88

Fig. 21. Rezultati zajedničkog Word2Vec - NamesFrequency - TF-IDF modela.

Model koji za značajke koristi samo frekvencije imena ima uspješnost od 53.70 %, a zajednički model ima uspješnost od 68.58 %, što je poboljšanje u odnosu na Word2Vec model ali i dalje je znatno manje uspješno nego TF-IDF modela. Testirali smo i model koji koristi sve

[5] https://scikit-learn.org/stable/

značajke (Word2Vec, TF-IDF i frekvenciju imena), taj model bio najbolji s uspješnosti od 74.82 %, što je znatno manje nego što su postigli prva tri tima na natjecanju.

B. BERT model

BERT (Bidirectional Encoder Representations from Transformers). Prvi put predstavljen u listopadu 2018., BERT je *word to vector* model za procesuiranje prirodnog jezika. Napravio ga je tim Googleovih znanstvenika s Jacobom Devlinom na čelu. Danas se koristi u Googleovom *search engine*-u za rangiranje stranica (*PageRank)* i istaknute isječke. BERT je baziran na transformeru, što znači da procesuira riječi u odnosu na ostale riječi u rečenici, umjesto da ih procesuira jednu po jednu kao prethodni *word to vector* modeli. Promotrimo sljedeću rečenicu:

"After stealing money from the bank vault, the bank robber was seen fishing on the Mississippi river bank."

U prethodnoj rečenici riječ "bank" pojavljuje se 3 puta, dvaput u kontekstu banke, a jednom u kontekstu obale rijeke. Kosinusna sličnost između vektora prve dvije pojave riječi "bank" je 0.94, dok je kosinusna sličnost između druge i treće pojave riječi "bank" 0.69.

Postoje dva tipa BERT modela, BERTbase i BERTlarge. BERTbase (BERTlarge) model koristi 12 (24) slojeva transformatorskog bloka sa skrivenom veličinom od 768 (1024) (broj značajki skrivenog stanja (hidden state) ZA RNN) i 12 (16) glava samopažnje te ima ~110M (~340M) parametara koji se mogu trenirati.

BERT uzima jednu ili dvije rečenice, koje se onda odvajaju separatorom. Svakoj riječi pridružen je jedan od ~30k tokena. Riječi koje nisu u vokabularu, BERT razdvaja na podriječi i zasebne znakove.

```
text = "Here is the sentence I want embeddings for."
marked_text = "[CLS] " + text + " [SEP]"

tokenized_text = tokenizer.tokenize(marked_text)

print (tokenized_text)
```

```
['[CLS]', 'here', 'is', 'the', 'sentence', 'i', 'want', 'em', '##bed', '##ding', '##s', 'for', '.', '[SEP]']
```

Fig. 22. Primjer tokeniziranja rečenice u BERT modelu.

BERT je originalno istreniran na engleskoj Wikipediji i na elektroničkoj kolekciji Američkih tekstova Brown Corpus. Budući da je jedan od ciljeva treniranja BERT-a bio next sentence prediction, odlučili smo se za njega pri rješavanju problema klasifikacije (*agree* / *discuss* / *disagree*).

Za našu klasifikaciju koristili smo već postojeću implementaciju iz 0. Korišten je predtreniran BERTbase modela iz Python biblioteke Transformers[6] te je podešen za *sentence entailment*. Model je treniran samo na povezanim primjerima (koji imaju odnos *agree*, *disagree* ili *discuss*) iz skupa podataka za treniranje. 13427 primjera podijeljeno je na 10742 (80%) primjera za treniranje i 2685 (20%) primjera za validaciju. Korišten je AdamW optimizator. Model je treniran u 10 epoha.

Zbog načina na koji je BERT implementiran i ograničenja BERT-a na 512 tokena, za klasifikaciju smo uspoređivali samo naslov s prvom rečenicom članka.

Izabrani model pokazao se uspješnijim u klasificiranju povezanih odnosa (agree, disagree ili discuss) od modela prva trim tima. Uzevši klasifikaciju related / unrelated od modela prva 3 tima, te korištenjem BERT modela za daljnju klasifikaciju related odnosa, dobili smo sljedeće rezultate:

- Uspješnost zajedničkog BERT modela i modela tima UCL je 83.43 %.
- Uspješnost zajedničkog BERT modela i modela tima Athene je 83.05 %.
- Uspješnost zajedničkog BERT modela i modela tima Talos je 83.62 %.

U tablicama na slikama Fig. 23, Fig. 24 i Fig. 25 prikazane su uspješnosti spomenutih modela.

	Agree	Disagree	Discuss	Unrelated	Accuracy (%)
Agree	976	201	612	114	51.29
Disagree	114	224	243	116	32.14
Discuss	382	237	3583	262	80.26
Unrelated	124	13	249	17963	97.90

Fig. 23. Rezultati zajedničkog BERT modela i modela tima UCL.

	Agree	Disagree	Discuss	Unrelated	Accuracy (%)
Agree	955	195	596	157	50.18
Disagree	115	210	223	149	30.13
Discuss	379	204	3531	350	79.10
Unrelated	36	2	100	18211	99.25

Fig. 24. Rezultati zajedničkog BERT modela i modela tima Athene.

[6] https://huggingface.co/docs/transformers/index

	Agree	Disagree	Discuss	Unrelated	Accuracy (%)
Agree	928	199	592	184	48.77
Disagree	114	229	239	115	32.86
Discuss	386	224	3620	234	81.09
Unrelated	41	0	197	18111	98.70

Fig. 25. Rezultati zajedničkog BERT modela i modela tima Talos.

V. ZAKLJUČAK

S obzirom na to da su prva tri tima na natjecanju imali visok postotak uspješnosti, naš jednostavni model koji za klasifikaciju koristi stabla odluke nije bio na njihovoj razini uspješnosti, ali pomoću BERT modela uspjeli smo nadograditi modele koje su koristili prva tri tima. Koristeći njihovu klasifikaciju za *unrelated* i *related* te našu klasifikaciju BERT modelom, bolje smo raspoznali odnose *agree*, *discuss* i *disagree* između naslova i sadržaja vijesti .

Zadatak za daljnju analizu bio bi istrenirati BERT model na većem broju rečenica iz sadržaja članka, što bi se moglo tako da se veći broj rečenica svrsta pod "drugu" rečenicu u tokenizaciji ili uspoređivanjem naslova sa svakom rečenicom teksta zasebno.

REFERENCES

[1] Fake News Challenge,
http://www.fakenewschallenge.org/.

[2] UCL Machine Reading - FNC-1 Submission,
https://github.com/uclnlp/fakenewschallenge.

[3] Athene system,
https://github.com/hanselowski/athene_system.

[4] Fake News Challenge - Team SOLAT IN THE SWEN,
https://github.com/Cisco-Talos/fnc-1.

[5] Benjamin Riedel, Isabelle Augenstein, Georgios P. Spithourakis, Sebastian Riedel,
A simple but tough-to-beat baseline for the Fake News Challenge stance detection task,
https://arxiv.org/abs/1707.03264.

[6] Andreas Hanselowski ,
Team Athene on the Fake News Challenge,
https://medium.com/@andre134679/team-athene-on-the-fake-news-challenge-28a5cf5e017b.

[7] Sean Baird, Doug Sibley, Yuxi Pan,
Talos Targets Disinformation with Fake News Challenge Victory
https://blog.talosintelligence.com/2017/06/talos-fake-news-challenge.html.

[8] Andreas Hanselowski, Avinesh PVS, Benjamin Schiller, Felix Caspelherr, Debanjan Chaudhuri, Christian M. Meyer, Iryna Gurevych,
A Retrospective Analysis of the Fake News Challenge Stance Detection Task,
https://arxiv.org/abs/1806.05180.

[9] Jakob Uszkoreit,
Transformer: A Novel Neural Network Architecture for Language Understanding,
https://ai.googleblog.com/2017/08/transformer-novel-neural-network.html.

[10] Rani Horev,
BERT Explained: State of the art language model for NLP,
https://towardsdatascience.com/bert-explained-state-of-the-art-language-model-for-nlp-f8b21a9b6270.

[11] Dharti Dhami,
Understanding BERT — Word Embeddings,
https://medium.com/@dhartidhami/understanding-bert-word-embeddings-7dc4d2ea54ca.

[12] Fine-tuning pre-trained transformer models for sentence entailment,
https://towardsdatascience.com/fine-tuning-pre-trained-transformer-models-for-sentence-entailment-d87caf9ec9db.

[13] Cosine similarity,
https://en.wikipedia.org/wiki/Cosine_similarity

[14] Term frequency–inverse document frequency
https://en.wikipedia.org/wiki/Tf%E2%80%93idf.

[15] Latent Dirichlet allocation,
https://en.wikipedia.org/wiki/Latent_Dirichlet_allocation.

[16] Latent semantic analysis,
https://en.wikipedia.org/wiki/Latent_semantic_analysis.

[17] n-gram,
https://en.wikipedia.org/wiki/N-gram.

[18] Singular value decomposition,
https://en.wikipedia.org/wiki/Singular_value_decomposition.

[19] Word2vec
https://en.wikipedia.org/wiki/Word2vec.

[20] Rectifier (neural networks),
https://en.wikipedia.org/wiki/Rectifier_(neural_networks)

[21] Softmax function.
https://en.wikipedia.org/wiki/Softmax_function

[22] Decision tree learning.
https://en.wikipedia.org/wiki/Decision_tree_learning

[23] Decision Trees Explained,
https://towardsdatascience.com/decision-trees-explained-3ec41632ceb6.