Fran Mišić, Andrej Slapničar, Iva Sokolaj, Roko Torbarina

Travanj, 2022.

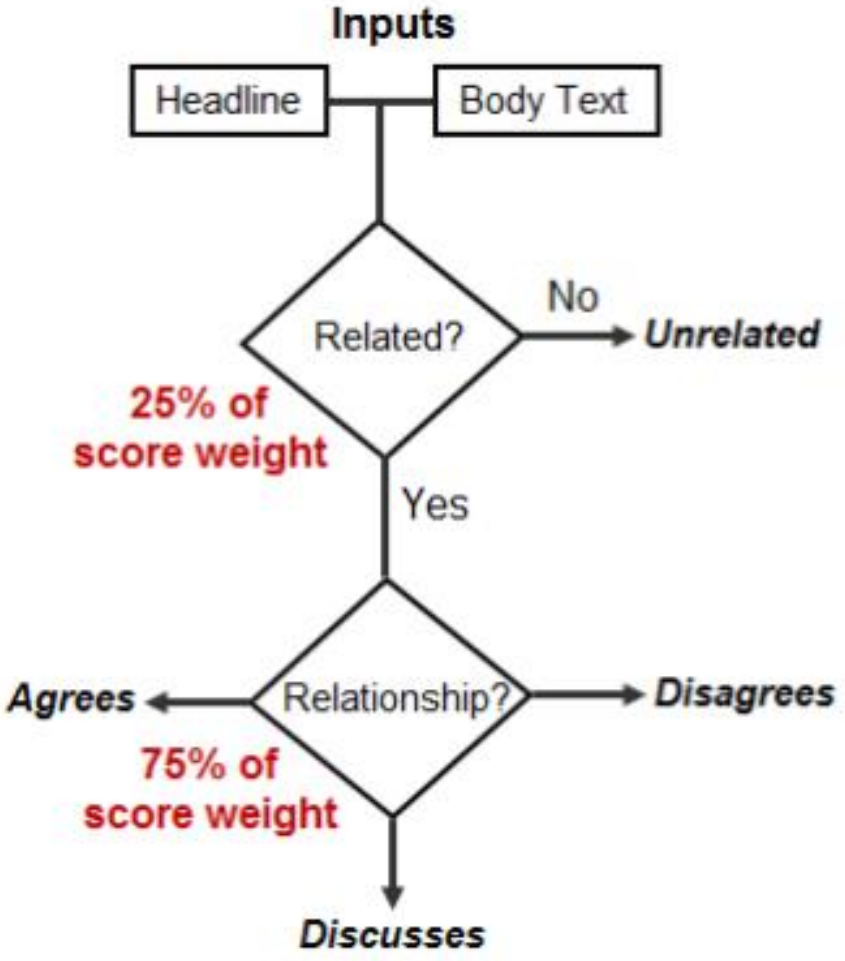
Opis problema

- Fake news challenge - FNC-1.
- Skup podataka za treniranje od 49972 instanci:
  - naslov vijesti,
  - sadržaj vijesti,
  - odnos naslova i sadržaja.
- Mogući odnos naslova i sadržaja:
  - slaganje (*agree*),
  - Neslaganje (*disagree*),
  - Diskusija (*discuss*),
  - Nepovezanost (*unrelated*).

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

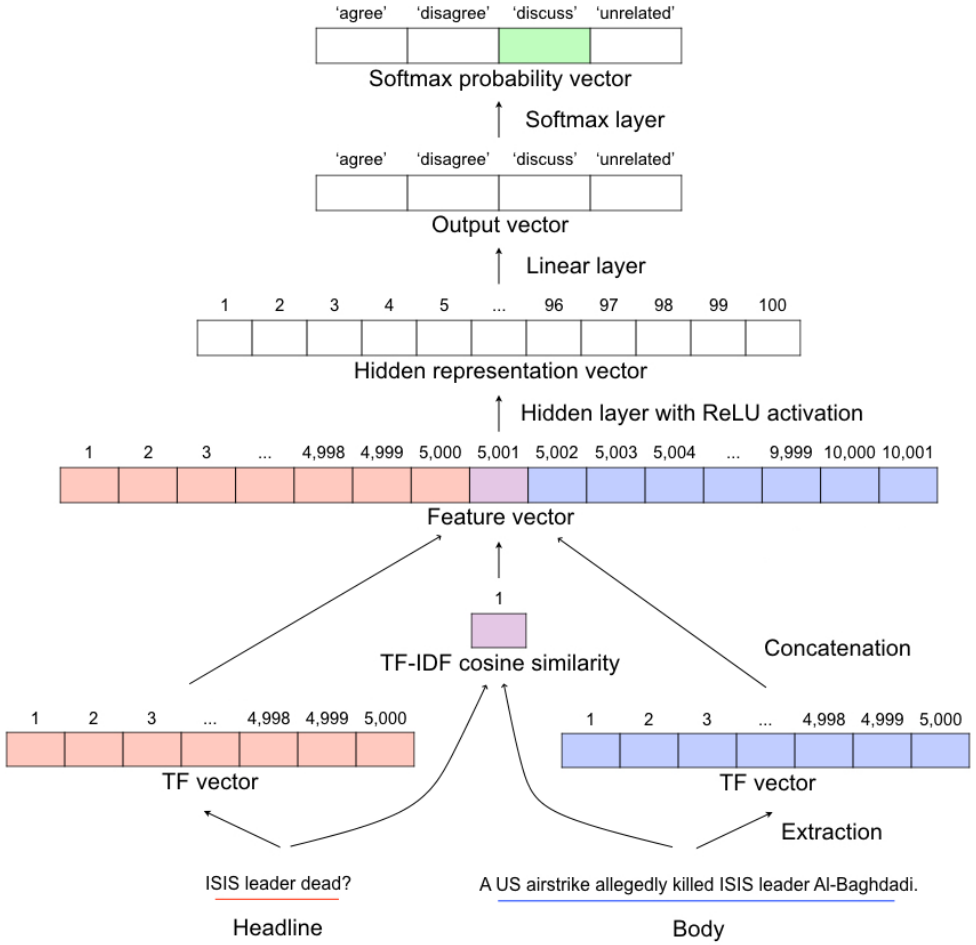
Opis problema

- Konačni rezultati bodovali su se prema idućim smjernicama:
  - +0.25 bodova ako je točno određena povezanost (unrelated /related),
  - +0.75 bodova ako je određena točna povezanost (agree / disagree / discuss).
- Skup podataka za testiranje od 25413 instanci (7064 bilo je povezano, a 18349 nepovezano).
- Maksimalan broj bodova bio je 7064 * 1 + 18349 * 0.25 = 11651.25.



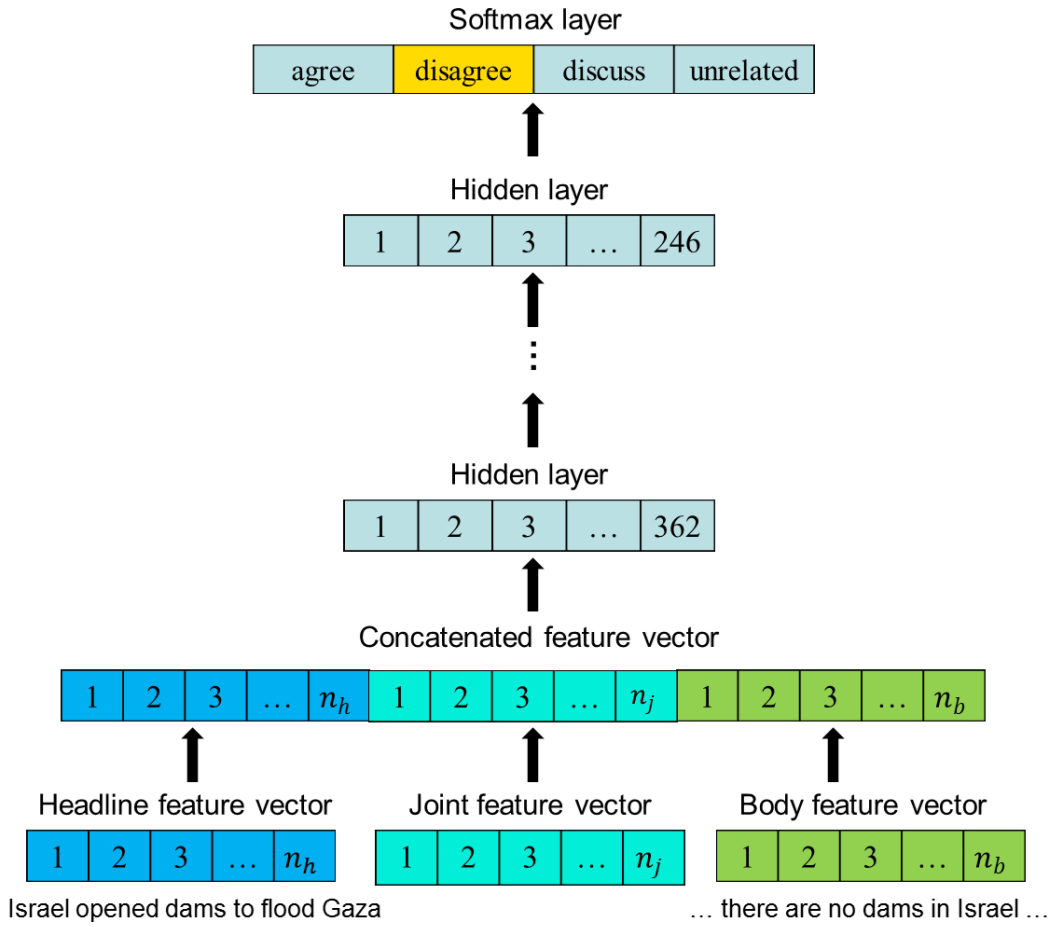
Najbolji timovi na natjecanju - UCL

- Treće mjesto na FNC-u osvojio je tim UCL Machine Reading.
- Korišten je MLP klasifikator s jednim skrivenim slojem od 100 jedinica i sa softmax-om na izlazu.
- Za nelinearnost skrivenog sloja iskorištena je ReLU aktivacijska funkcija
- Značajke:
  - vektori frekvencija 5000 najčešćih riječi
  - kosinusna sličnost TF-IDF vektora



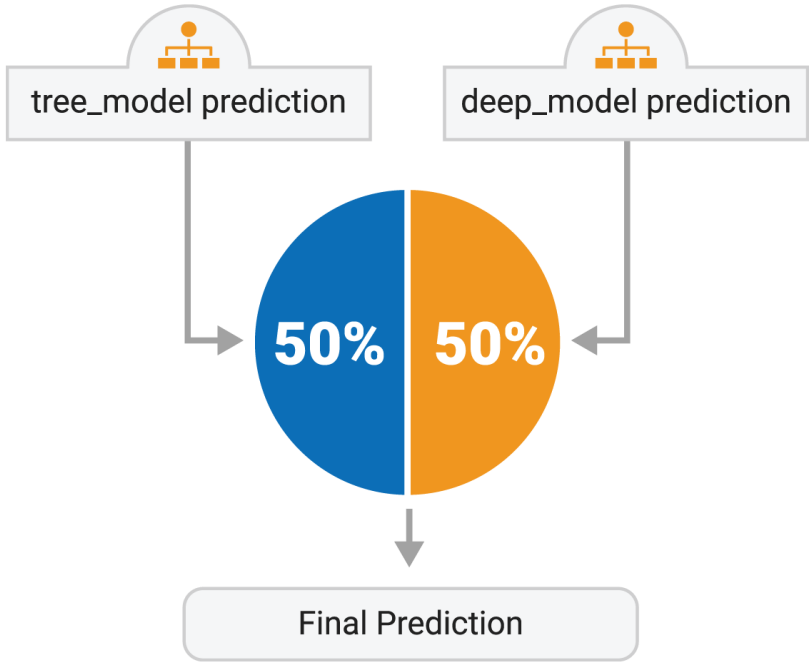
Najbolji timovi na natjecanju - Athene

- Drugo mjesto na FNC-u osvojio je tim Athene (UKP Lab).
- Korišten je MLP klasifikator s 6 skrivenih slojeva i softmax slojem.
- Značajke:
  - kosinusna sličnost,
  - latentna Dirichletova alokacija (LDA),
  - latentno semantičko indeksiranje (LSI).



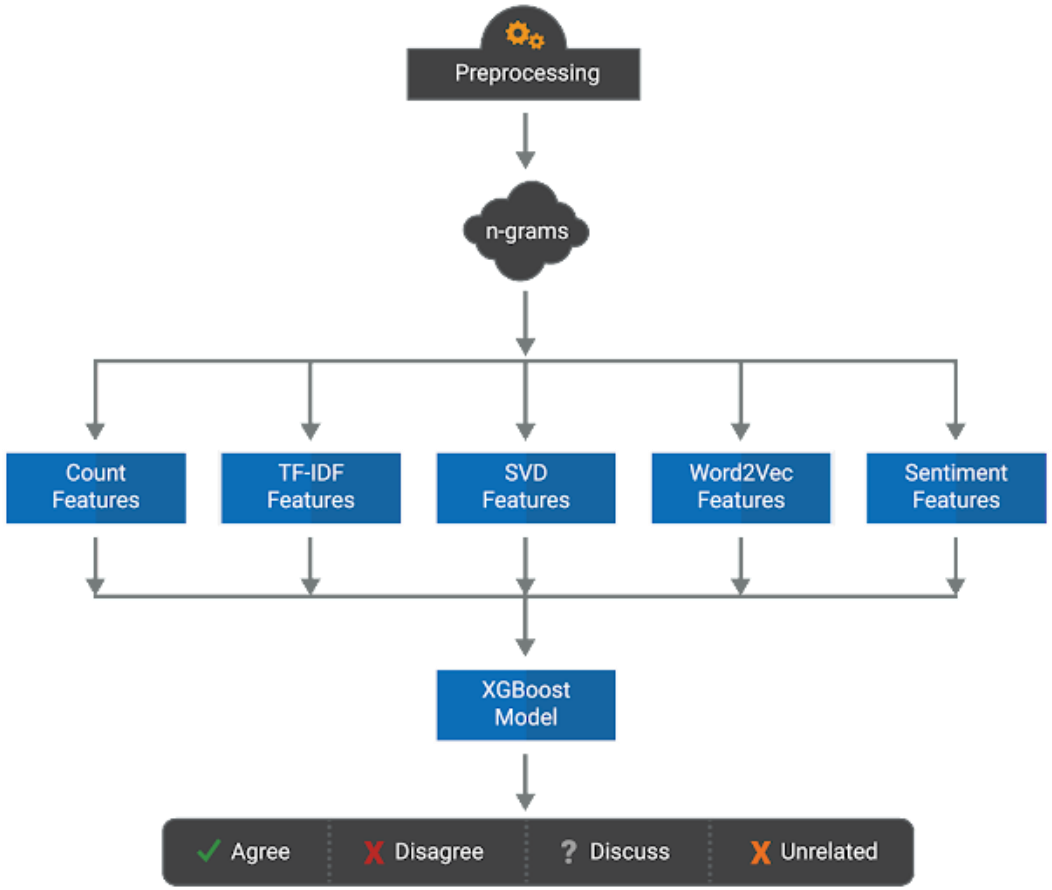
Najbolji timovi na natjecanju - Talos

- Prvo mjesto na FNC-u osvojio je tim SOLAT in the SWEN (Talos) .
- Njihov model rezultate donosi na temelju aritmetičke sredine predikcija dobiveni:
  - metodom stabla odlučivanja s pojačanim gradijentom (GBDT),
  - dubokim konvolucijskim neuronskim mrežama (deep CNN).



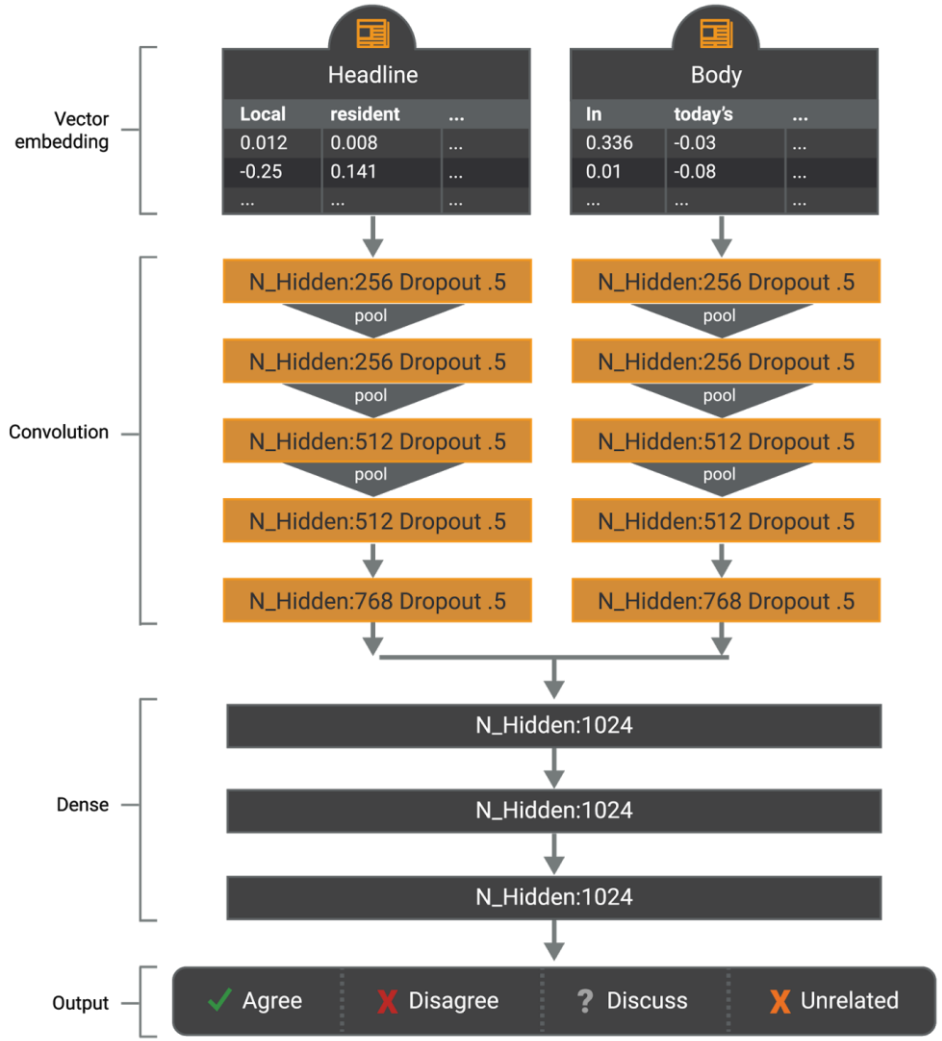
Najbolji timovi na natjecanju - Talos

- Model stabla odlučivanja s pojačanim gradijentom (GBDT):
  - Značajke:
    - frekvencije unigrama, bigrama i trigrama,
    - singularna dekompozicija (SVD) TF–IDF vektora,
    - Word2Vec *embeddings*,
    - sentiment.
  - Korištena je XGBoost implementacija ručno podešena za navedene značajke vijesti.



Najbolji timovi na natjecanju - Talos

- Model s dubokim konvolucijskim neuronskim mrežama (deep CNN):
  - Koristi se jednodimenzionalnu konvolucijsku mrežu (1D-CNN) na naslovu i sadržaju vijesti.
  - Zatim se na izlazu koji daje 1D-CNN trenira višeslojni perceptron (MLP)
  - Značajke:
    - Word2Vec *embeddings*.



Najbolji timovi na natjecanju - uspješnost

- Prvu nagradu od USD 1000 na natjecanju osvojio je tim SOLAT in the SWEN (Talos) s ukupnih 9556.50 bodova, što je 82.02 % maksimalnog broja bodova.

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

Najbolji timovi na natjecanju - uspješnost

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

	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

- Rezultati tima Athene na test *dataset*-u iz natjecanja

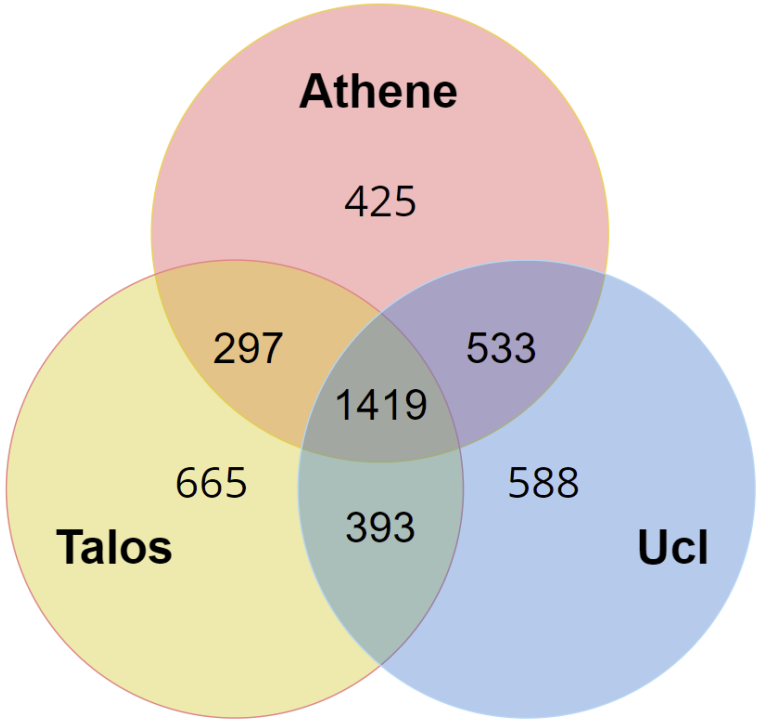
	Agree	Disagree	Discuss	Unrelated	Accuracy (%)
Agree	851	69	826	157	44.72
Disagree	241	66	241	149	9.47
Discuss	466	37	3611	350	80.89
Unrelated	19	4	115	18211	99.25

- Rezultati tima Talos 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

Najbolji timovi na natjecanju - greške

- Modeli rade slične greške.
- Sva tri tima najslabiji uspjeh imali su s *disagree* klasifikacijom zbog:
  - mali broj instanci s odnosom *disagree*
  - korištene značajke nisu pogodne za određivanje razlika između *agree*, *disagree* i *discuss*.



Najbolji timovi na natjecanju - greške

- Najčešći razlozi pogrešnih klasifikacija:
  - Podudaranje riječi između naslova i članka može nepovezane parove klasificirati kao povezane,

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

Najbolji timovi na natjecanju - greške

- Najčešći razlozi pogrešnih klasifikacija:
  - Korištenje dva sinonima istog pojma u naslovu i sadržaju može povezane parove klasificirati kao nepovezane,

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

Najbolji timovi na natjecanju - greške

- Najčešći razlozi pogrešnih klasifikacija:
  - Pojavljivanje određenih riječi kao što su “allegedly“, “according to“ i “said“, u sadržaju može rezultirati pogrešnom discuss klasifikacijom.

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

Naši modeli - klasifikacija koristeći stabla odlučivanja

- Jednostavni model koji za klasifikaciju koristi stabla odluke.
- Značajke:
  - kosinusna sličnost aritmetičke sredine Word2Vec *embeddings*-a
  - kosinusna sličnost TF-IDF vektora
  - zbroj frekvencija imena iz naslova u sadržaju vijesti
- Korištena je implementacija klasifikatora sa stablima odlučivanja iz Python biblioteke Scikit-learn.
- Uspješnost modela:
  - Word2Vec model -> 65.95 %,
  - TF-IDF model -> 73.03 %,
  - Word2Vec - TF-IDF model -> 74.27 %,
  - NamesFrequency model -> 53.70 %,
  - Word2Vec - NamesFrequency model -> 68.58 %,
  - Word2Vec - TF-IDF - NamesFrequency model -> 74.82 %

Naši modeli - klasifikacija koristeći stabla odlučivanja

- Rezultati Word2Vec modela.

	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

- Rezultati TF-IDF 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

- Rezultati zajedničkog Word2Vec - 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

Naši modeli - klasifikacija koristeći stabla odlučivanja

- Rezultati NamesFrequency modela.

	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

- Rezultati zajedničkog Word2Vec - 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

- Rezultati zajedničkog Word2Vec - NamesFrequency - TF-IDF 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

Naši modeli - BERT model

- BERT (*Bidirectional Encoder Representations from Transformers*) je Google-ov *word to vector* model za procesuiranje prirodnog jezika.
- BERT je baziran na transformeru, što znači da procesuira riječi u odnosu na ostale riječi u rečenici.
- Primjer:
  - “After stealing money from the bank vault, the bank robber was seen fishing on the Mississippi river bank.“
  - 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.

```
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]']
```

Naši modeli - BERT model

- 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*).
- Korišten je predtreniran BERTbase modela iz Python biblioteke Transformers.
- Korišten je AdamW optimizator. Model je treniran u 10 epoha.
- Za klasifikaciju smo uspoređivali samo naslov s prvom rečenicom članka.
- Uspješnost modela:
  - zajednički BERT model i modela tima UCL -> 83.43 %,
  - zajednički BERT model i modela tima Athene -> 83.05 %,
  - zajednički BERT model i modela tima Talos -> 83.62 %.

Naši modeli - BERT model

- Rezultati zajedničkog BERT modela i modela tima UCL.

	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

- Rezultati zajedničkog BERT modela i modela tima Athene.

	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

- Rezultati zajedničkog BERT modela i modela tima Talos.

	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

Zaključak

- Lošija uspješnost model koji za klasifikaciju koristi stabla odluke.
- 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.


hero
vote
rumor
Like
truth
trust
hack
fraud
Share
terror
war
internet
election
secret
web
hate
protest
Fake
News
Challenge
videos
traitor
trigger
panic
mob
fall
say
source
true
cyber
group
party
site
kill
leader
argue
danger
ugly
crowd
enemy
foreign
politics
false
lies
law
world
fight
urgent
media
power
photos
fear
risk
speech
dictator
force
shout
unqualified
opponent
democracy