# DHH23 – Helsinki Digital Humanities Hackathon 2023

*David BORDON*
Univerza v Ljubljani, Filozofska fakulteta

Maja 2023 je v Helsinkih potekal humanistični *hackathon*.

Ob omembi besede *hackathon* se marsikomu pred očmi izriše slabo prezračevan kletni prostor, kamor se čez konec tedna – namesto da bi preživeli nekaj časa na svežem zraku, občasno oplazili kakšno zelenico in nadomestili kronično pomanjkanje vitamina D – zapre nekaj programerjev ali razvijalcev videoiger in hudo neprespani po četrtem obroku, ki ga predstavlja pica, kot vse obroke prej, skuša ugotoviti, zakaj je VS Code tako zelo poln klicajev. Klasičen primer, ko si preobremenjeni ljudje vzamejo odmor od »cruncha«[1] z *drugačnim* »crunchem«.

V Helsinkih je potekal, kot omenjeno, humanistični hackathon, kar je nekoliko drugačna zadeva. Klet je bila prostornejša, svetlejša in dobro prezračevana, izbor hrane je bil bogat, neprespanost je bila posledica posameznikove neodgovornosti, delo je potekalo ob rednem delovnem času, konec tedna pa je bil popolnoma prost. Ukvarjali smo se s tem, da smo na nekoliko bolj igriv način izvedli znanstveno raziskavo, ki je povezovala področji humanistike in računalništva, končni rezultat pa je bil konferenčni poster, ki smo ga predstavili na zaključnem dogodku.

1 Obdobje zelo intenzivnega dela, ki se nikoli ne konča.

*Bordon, D.: DHH23 – Helsinki Digital Humanities Hackathon 2023. Slovenščina 2.0, 11(2): 92–101.*

*1.19 Recenzija, prikaz knjige, kritika / Review, book review, critique*

DOI: https://doi.org/10.4312/slo2.0.2023.2.92-101

Po tem splošnem orisu se lahko posvetimo temu, kaj Helsinki Digital Humanities Hackathon sploh je. Dogodek ima že močno tradicijo – prvič so ga priredili leta 2015, od tedaj (z izjemo 2020) poteka letno pod okriljem Univerze v Helsinkih in v praksi demonstrira njihov pristop k poučevanju digitalne humanistike, kjer je poudarek na posameznikih in njihovih veščinah, obenem pa gre za poligon, kjer lahko ti posamezniki razvijejo skupni jezik in plodno sodelujejo na skupnem projektu. Hackathon soorganizira Univerza Aalto, dodatno podporo pa nudita konzorcija CLARIN in DARIAH, ki sta denimo letos krila stroške potovanja in namestitve za 20 udeležencev.

Aktivnosti potekajo v »kleti« centra Oppimiskeskus Minerva (Učni center Minerva), v kateri je poleg skupne dvorane z delovnimi mizami na voljo še pet dobro opremljenih zastekljenih »učilnic«, kjer lahko skupine delajo na svojih projektih. Zgoraj, v pritličju, je menza, s katero upravlja lokalna študentska organizacija – iz pogovora z enim od koordinatorjev naše ekipe sem izvedel (upam, da nisem česa razumel narobe), da se študentska organizacija helsinške univerze zelo dobro hrani z oddajanjem zemljišč ob glavni železniški postaji, kjer bi (preden je zrasla postaja) univerza morala imeti svoj kampus, sedaj pa so tam trgovine in poslovalnice multinacionalk, ki plačujejo neskromno najemnino. Dobiček gre tudi v sistem subvencionirane prehrane za študente, koncept, ki si ga na domači grudi, kjer imajo akterji tipa ŠOU povsem drugačne apetite, zelo težko predstavljamo.

Subvencije za prehrano so veljale tudi za vse udeležence dogodka, poleg tega pa so nam bile na voljo neomejene količine kave ali čaja – še več – vzporedno je potekalo tudi tekmovanje, kdo bo konzumiral največ omenjenih stimulantov, na koncu pa razglasitev obeh zmagovalcev, nevrotično tresočega Italijana in temperamentnega Grka, ki sta na glavo spila okrog 5 skodelic dnevno.

Hackathon je potekal 10 dni, od tega sta bila dva dneva, sobota in nedelja, popolnoma prosta. Nekateri smo ta čas izkoristili za dnevno potepanje po Helsinkih in obisk estonskega Tallina, ki je zgolj prijetno uro plovbe stran. Vsekakor, če je možnost, priporočam, da si bodoči udeleženci po dogodku vzamejo še kak dodaten dan dopusta, saj je mesto čudovito in kulturna ponudba izvrstna.

Namen dogodka je spodbujanje interdisciplinarnega sodelovanja med humanisti in računalničarji, o katerem je eden izmed organizatorjev, Eetu Mäkelä, govoril na vabljenem predavanju konference JTDH 2022[2]. Obenem je dogodek leta 2021 zaradi oteženih pogojev potovanja potekal na spletu, organizatorji pa so s tem pridobili ogromno povratnih informacij, kako dogodek optimalno zastaviti v bodoče. O tem je na TwinTalks delavnici na konferenci DH23 Graz govoril še en organizator, Mikko Tolonen, čigar prispevek si lahko ogledate na YouTube kanalu CLARIN ERIC[3].

Formalno se je hackathona udeležilo približno 60 oseb, ki so bile razdeljene v štiri različne skupine. Vsaka skupina je dobila podatkovno zbirko, na podlagi katere so osnovali znanstveno raziskavo o določeni temi. Te so bile sledeče:

- Metapodatki pisem in razumevanje družbe
- Interakcijska dinamika spletnega diskurza
- Zgodnje novoveške znanstvene publikacije
- Politična polarizacija v parlamentu

Slovenijo smo, poleg pisca tega prispevka, predstavljali še Katja Meden (INZ), Vid Klopčič (FRI), ukrajinska sodelavka na INZ Anna Kryvenko in v vlogi enega izmed vodij naše tematske skupine makedonski kolega Bojan Evkoski, takrat še raziskovalec na INZ – vsi smo bili del parlamentarne skupine, ki je raziskavo oblikovala okrog podatkovne zbirke ParlaMint.

Pred samim dogodkom smo imeli dve uvodni srečanji na spletu, kjer smo se spoznali med seboj, s podatkovno zbirko, postavili okvirna raziskovalna vprašanja in pretehtali metodološke pristope. V vmesnem času je računalniški del skupine že izvajal predprocesiranje podatkov, humanisti pa so zbirali gradivo za potencialni teoretski okvir.

Dogodek je bil izvrstno organiziran. Prvi dan so nas odgovorni seznanili s pogoji dela, roki za predstavitev napredka in vmesnih rezultatov, stranskimi projekti, družabnimi dogodki in večkrat poudarili, na koga se lahko obrnemo, če se čutimo v stiski zaradi kakršnih koli razlogov. Za tem smo pričeli z delom, ki je potekalo neprekinjeno vse

2 https://www.sdjt.si/wp/dogodki/konference/jtdh-2022/zbornik/#vabljena
3 https://youtu.be/kTjahw2q_5g?si=SmbgPEA0-z4w58Xe&t=9223

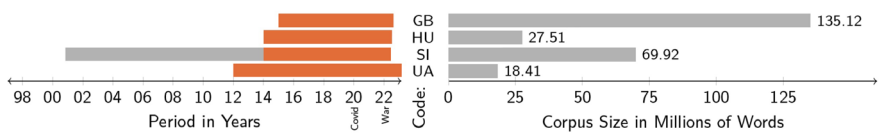
do predstavitve posterjev. Vzporedno smo imeli še nekaj »stranskih projektov«, kjer smo na nekoliko bolj eksperimentalen in umetniški način s popolnoma prostimi rokami uporabljali generativno umetno inteligenco za različne namene (recimo, multimedialno prevajanje poezije). Poleg dela v skupinah je bilo nekaj segmentov vezanih na sodelovanje med skupinami, denimo, letos smo vzpostavili kanal na družbenih omrežjih, s katerim upravljajo sodelujoči in ga bo v bodoče možno predajati naslednjim generacijam.

Sodelovanje s člani drugih skupin je bilo omogočeno v skupnem centralnem prostoru, kjer so bili delovni otoki. Najbolj plodne debate so bile na sporedu po koncu delovnega dneva, ko smo se zbrali tam, kjer je *žejni učenjak* doma – prostor, kjer smo tudi poglobili obstoječe vezi in sklenili nove.

Če se sedaj posvetim dejanskemu delu parlamentarne skupine – ukvarjali smo se z vprašanjem polarizacije v parlamentih. Naša raziskava se je osredotočila na odgovore na ključna vprašanja, povezana s polarizacijo, kot so:

1) Kako je mogoče meriti polarizacijo z računalniškimi metodami?
2) Kako se polarizacija kaže skozi čas?
3) Kako določene teme polarizirajo različne parlamente?

Za merjenje polarizacije, opazovanje njenega izražanja skozi čas in analizo, kako določene teme polarizirajo parlamente, smo uporabili računalniške metode in podatkovno zbirko ParlaMint 3.0 (različica beta), ki jo je predhodno izdal CLARIN ERIC. Gre za primerljiv večjezični korpus parlamentarnih razprav. Izbrali smo štiri države: **Združeno kraljestvo**, **Madžarsko**, **Ukrajino** in **Slovenijo** v spodaj navedenih obdobjih (Slika 1).




**Slika 1:** Levo: časovni obseg korpusa besedil, ki smo jih vključili v raziskavo (oranžno). Desno: velikost korpusov v milijonih besed. Vertikalno v sredini: država.

Naša metodologija je bila razdeljena na tri glavne dele:

1) Tematske podkorpuse
2) Numerične reprezentacije govorov
3) Merjenje polarizacije

Za izbiro specifičnih tem iz parlamentarnih govorov smo uporabili multidisciplinarni pristop, pri čemer smo izločili ključne besede za cepitev korpusov ter uporabili vektorsko vložitev ključnih besed po metodi LDA. Določili smo tudi pomembne leme in besedne zveze za vsako temo, pri čemer smo upoštevali generičnost in metaforično rabo. Prispevke in njihov vpliv na polarizacijo smo analizirali s filtriranjem govorov na podlagi teh ključnih besed ter jih ovrednotili s pomočjo »matrike zmede« (confusion matrix) in števila govorov (Tabela 1).

**Tabela 1:** *Število govorov za določeno podtemo, pridobljenih s ključnimi besedami, ki so bile vključene v analizo*

| Tema | SI | GB | HU | UA |
|---|---|---|---|---|
| **EU** | 13943 (11 %) | 38938 (10 %) | 5933 (14 %) | 5863 (8,4 %) |
| **vojna** | 8686 (7,0 %) | 15039 (4.0 %) | 2141 (5.0 %) | 7101 (10 %) |
| **zdravstvo** | 13802 (11 %) | 45022 (12 %) | 4583 (11 %) | 4546 (6.5 %) |

*Opomba. V oklepaju: procent pojavnosti v celotnem korpusu.*

Naš cilj je bil ustvariti primerljive reprezentacije večjezičnih govorov. Uporabili smo model SBERT za vektorske vložitve govorov, kar nam je omogočilo generiranje visokodimenzionalnih vektorjev za optimizirane semantične primerjave. Poleg tega smo z ekstrakcijo sentimenta s pomočjo regresije na podlagi modela RoBERTa določili vrednosti sentimenta za vsak govor.


**Slika 2:** Ocena sentimenta za štirimesečja. Oranžna – pozitiven sentiment; modra – negativen sentiment.

Za vrednotenje polarizacije smo uporabili štiri tehnike, in sicer vložitve razlik med nasprotnimi strankarskimi skupinami, analizo sentimenta in primerjavo nasprotnih poudarkov na podtemah BERTopic. Najprej smo preučili razlike v vložitvah govora (vektorjih) med nasprotnimi strankarskimi skupinami. To nam je pomagalo razumeti, kako različne stranke izražajo svoja stališča in obseg polarizacije med njimi. Izvedli smo tudi analizo sentimenta, da bi ugotovili splošno razpoloženje političnih strank. S primerjavo nasprotujočih si stališč o določenih podtemah smo dobili vpogled v njihovo polarizacijsko naravo. Nazadnje smo izvedli diahrono analizo, da bi preučili, kako se je polarizacija spreminjala skozi čas, in sicer z analizo sentimenta in vložitev.

Naše delo je privedlo do več pomembnih ugotovitev:

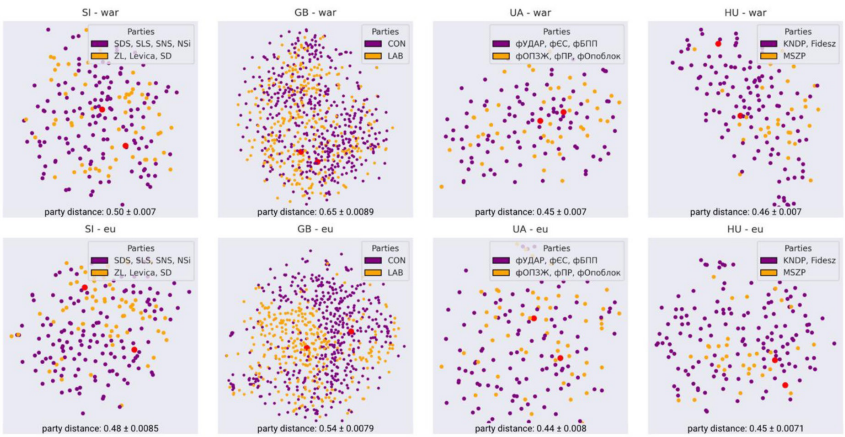
- **Grozdenje tem**: Z uporabo tehnike BERTopic smo iz krovnih tem uspešno izluščili podteme. To nam je omogočilo, da smo se poglobili v govore in primerjali usmeritve zgodovinsko nasprotujočih si strank v posameznih državah. S prepoznavanjem in združevanjem teh podtem smo celovito razumeli vprašanja, ki prispevajo k polarizaciji parlamentarnih razprav (Tabela 2).

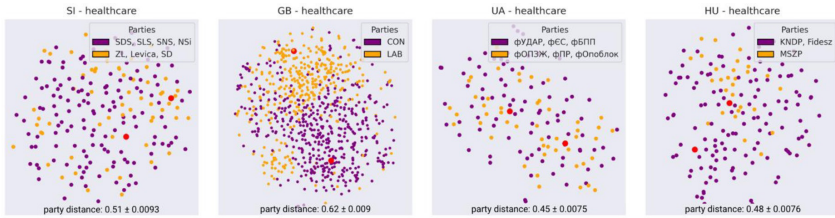
**Tabela 2:** *Izbrane podteme BERTopic, ki poudarjajo razkol med nasprotujočimi si političnimi skupinami*

| Great Britain | | | |
|---|---|---|---|
| Theme | Topic | Focus % (CON / LAB) | Sentiment (CON / LAB) |
| EU | Brexit Referendum | 11.1 / 17.6 | 0.10 / -0.08 |
| War | Ukraine-Russia War | 13.4 / 10.6 | -0 08 / -0 30 |
| Healthcare | Covid | 30.1 / 22.5 | 0.17 / -0.35 |
| Ukraine | | | |
| | | Focus% (Pro-UA* / Pro-RU*) | Sentiment (Pro-UA* / Pro-RU*) |
| EU | Language Policy | 10.30 / 47.50 | -0.34 / -0.55 |
| War | Legislations in War | 9.60 / 22.40 | -0.41 / -0.58 |
| Healthcare | Organ Transplantation | 16.80 / 1.10 | 0.03 / -0 88 |

| Hungary | | | |
|---|---|---|---|
| | | Focus % (Fidesz-KDNP / MSZP) | Sentiment (Fidesz-KDNP / MSZP) |
| EU | Corruption Charges | 16.70 / 30.9 | -0.38 / -0.48 |
| War | Constitution Defense | 13.4 / 29.5 | -0.32 / -0.52 |
| Healthcare | Covid | 24.3 / 9.5 | -0 13 / -0 52 |
| **Slovenia** | | | |
| | | Focus % (SDS / SD) | Sentiment (SDS / SD) |
| EU | Tax Coffers | 6.90 / 10.8 | -0 04 / 0.02 |
| War | Veteran Pensions | 2.2 / 5.0 | -0 30 / -0 32 |
| Healthcare | Healthcare System | 14.2 / 8.0 | -0 26 / -0 48 |

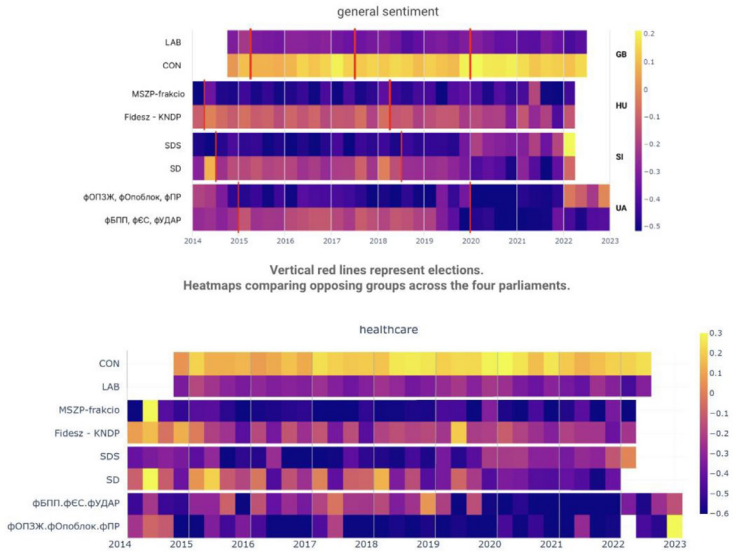
- **Osi polarizacije:** Z zmanjšanjem dimenzionalnosti in izrisom t-SNE smo vizualizirali odnose med govorci iz različnih strank znotraj vsake države in posamezne teme. Ta tehnika vizualizacije je razkrila jasne vzorce in delitve, zlasti v Veliki Britaniji, ko smo vektorje z visoko dimenzionalnostjo pretvorili v dvodimenzionalne vložitve; prepoznali smo osi polarizacije, ki kažejo, v kolikšni meri so bile stranke pri določenih vprašanjih razdeljene. Poleg tega je analiza sentimenta razkrila polarizacijo med prevladujočimi strankami, pri čemer je bil pozitiven sentiment večinoma opažen v koalicijah, negativen pa v opozicijskih skupinah. To je pokazalo velike razlike v pogledih in stališčih med političnimi skupinami, ki smo jih vključili v raziskavo (Slika 3).






**Slika 3:** Osi polarizacije.

- **Sentiment političnih strank:** V naši raziskavi smo poleg ugotavljanja polarizacije na širši ravni raziskali tudi polarizacijo čustev pri posameznih temah. Ugotovili smo, da so različna vprašanja, vključno z zdravstvom, vojno in Evropsko unijo, pokazala pomembno polarizacijo sentimenta med političnimi strankami. S to ugotovitvijo smo potrdili kontroverznost tem ter različna stališča in mnenja, ki jih imajo različne stranke. Razumevanje čustev, ki se povezujejo s temi temami, je ključno za razumevanje globine polarizacije in njenega vpliva na parlamentarni diskurz (Slika 4).[4]




**Slika 4:** Analiza sentimenta. Primerjava dveh nasprotujočih si strank iz vključenih parlamentov.

4 Omejitev odgovornosti in obvestilo o avtorstvu: Vsebino tehničnega dela na straneh 96–99 je kolektivno spisala skupina za potencialno objavljanje na blogih ali v poročilih, vodilno vlogo pri pisanju je imela kolegica Nikoletta Jablonczay. Avtor prispevka sem izvirno besedilo prevedel v slovenščino in ga rahlo prestrukturiral.

V sklopu raziskave smo naleteli na nekatere izzive, ki so nam ponudili priložnosti za nadaljnje raziskovanje:

- **Omejitve podatkov**: Različni obsegi podkorpusov so omejili možnost posploševanja modelov sentimenta in ovirali ponovljivost tematskega modeliranja.
- **Metodološke razlike**: Čeprav so pristopi, ki temeljijo na LLM (velikih jezikovnih modelih), ponudili dragocen vpogled, so razlike v razlagi tem predstavljale izziv za kvalitativno analizo.
- **Kvalitativna analiza**: Krepitev okvira kvalitativne analize po opravljeni raziskavi bi povečala zanesljivost naših metod.

Možnosti za nadaljnje raziskave vključujejo razvoj konkretne ontologije polarizacije za pomoč pri izbiri in prepoznavanju lastnosti, standardizacijo računalniških metodologij ter vključitev razumljivih metod umetne inteligence in »bele škatle« za boljšo interpretacijo, večjo transparentnost in naknadno analizo.

Raziskavo smo predstavili v obliki konferenčnega posterja[5], na spletu pa so na voljo tudi posterji ostalih treh skupin[6, 7, 8].

Za zaključek pa nekaj splošnih nasvetov za bodoče udeležence – Skandinavija je, za slovenski standard, draga, predvsem kar se tiče nočitev. Priporočljiv je skupen najem nastanitve, kar so spodbujali tudi organizatorji s predčasno vzpostavitvijo temu namenjenega kanala na projektnem Slacku. Stroški prehranjevanja niso visoki, predvsem zaradi odlične menze v objektu Minerva, nestandardno dragi so samo potencialni *promilni priboljški*, za katere je v povprečju treba odšteti dvokratnik slovenske cene, je pa stranski učinek intelektualnega doprinosa ob spremljajočem druženju neprecenljiv. Helsinki so razmeroma majhna prestolnica, ki je obvladljiva peš, javni promet je izvrsten, na voljo so tramvaji in podzemna železnica, vozovnice pa je ob vstopu na postajo podzemne ali na tramvaj moč kupiti prek aplikacije HSL. Ob odsotnosti direktnih povezav je Helsinško letališče zelo dobro povezano z evropskimi vozlišči, karte pa ne bi smele biti predrage, če se potovanje načrtuje dovolj zgodaj.

5 https://www.helsinki.fi/assets/drupal/2023-06/dhh23-parliament-poster.pdf
6 https://www.helsinki.fi/assets/drupal/2023-06/dhh23-disc-poster-comp1.pdf
7 https://www.helsinki.fi/assets/drupal/2023-06/dhh23-letters-poster.pdf
8 https://www.helsinki.fi/assets/drupal/2023-06/dhh23-earlymodern-poster.pdf

In še čisto za konec – Helsinki Digital Humanities Hackathon je izvrsten dogodek, kjer posameznik spozna, da je delo v skupini, ki jo sestavljajo osebe iz drugačnih akademskih ozadij, ki imajo povsem drugačne pristope in tradicije, vznemirljivo. Obenem je ta pogled onstran svojega vrtička v zgodnji fazi raziskovalnega udejstvovanja izjemnega pomena, saj pri posamezniku pusti ihtečo željo po novih oblikah sodelovanja, in ne zapiranja v varnost lastnega balončka. Vsem, ki imate možnost, da se dogodka udeležite, svetujem, da spremljate relevantne spletne strani[9] in v prihajajočih letih izkusite, kaj vse vam lahko da.

9 https://www.helsinki.fi/en/digital-humanities