Klasifikacija samomorilskih pisem

Dejan Rupnik
Fakulteta za elektrotehniko, računalništvo in informatiko
Koroška cesta 46
Maribor, Slovenija
dejan.rupnik@student.um.si

Denis Ekart
Fakulteta za elektrotehniko, računalništvo in informatiko
Koroška cesta 46
Maribor, Slovenija
denis.ekart@student.um.si

Gregor Kovačevič
Fakulteta za elektrotehniko, računalništvo in informatiko
Koroška cesta 46
Maribor, Slovenija
gregor.kovacevic@student.um.si

Dejan Orter
Fakulteta za elektrotehniko, računalništvo in informatiko
Koroška cesta 46
Maribor, Slovenija
dejan.orter@student.um.si

Alen Verk
Fakulteta za elektrotehniko, računalništvo in informatiko
Koroška cesta 46
Maribor, Slovenija
alen.verk@student.um.si

Borko Bošković
Fakulteta za elektrotehniko, računalništvo in informatiko
Koroška cesta 46
Maribor, Slovenija
borko.boskovic@um.si

POVZETEK

V našem delu smo se osredotočili na klasifikacijo poslovilnih pisem samomorilcev. Pozornost smo posvetili na razpoznavo pristnih pisem te narave, od pisem, ki to niso, oz. so le-ta lažna. S pomočjo procesiranja naravnega jezika in algoritmov strojnega učenja želimo doseči, da se pristna pisma v večini ločijo od lažnih. Implementirali smo program v programskem jeziku Python, pripravili korpus in izvedli nadzorovano strojno učenje. Pisma smo klasificirali z metodami: DecisionTreeClassifier, SVC, GaussianProcessClassifier, AdaBoostClassifier, KNeighborsClassifier, RandomForestClassifier, MLPClassifier, GaussianNB in QuadraticDiscriminantAnalysis. Najboljše rezultate smo dosegli z odločitvenim drevesom, kjer smo dosegli 68% natančnost.

KLJUČNE BESEDE

klasifikacija samomorilskih pisem, procesiranje naravnega jezika, odločitvena drevesa

1. UVOD

Zaradi samomora vsako leto umre več kot 800.000 oseb (v Sloveniji več kot 300), približno 25-krat toliko pa jih samomor poskuša narediti [14][11]. Velikokrat, ko vidimo samomorilen zapis, smo v dilemi ali ta oseba misli resno ali pa je to poizkus iskanja pozornosti [7]. Nekatera pisma so tudi lažna - na primer pri umorih, kjer bi želel storilec prikazati, da je žrtev storila samomor. Posledično je njihova klasifikacija pomembna za preventivo ter za razreševanje morebitnih nejasnosti.

Za samo izvedbo klasifikacije smo najprej potrebovali pristna pisma samomorilcev ter lažna pisma. Korpus pristnih pisem smo pridobili s spletnih strani [12, 3, 6, 4], lažna pa smo sestavili sami. Vsa pisma so v angleškem jeziku. Uporabili smo 63 pristnih in 19 lažnih pisem. Postopek analize pisem je potekal v več korakih. Prvi izmed njih je bilo predprocesiranje besedila, kjer smo kot rezultat dobili prečiščeno besedilo (tj. samo standardni znaki, brez simbolov in ločil). Nato smo izvedli oblikoslovno označevanje besed in povezavo ključnih besed s posameznimi koncepti iz tega področja. Naredili smo statistiko posameznih pisem (povprečno število besed ipd.). Izvedli smo tudi teste berljivosti, kateri so izračunali dve različni metriki. Tudi te smo uporabili pri strojnem učenju. Dodatno smo opravili tudi analizo čustev, nato pa vse rezultate združili v datoteke CSV. Te so bile osnova za gradnjo odločitvenih dreves pri strojnem učenju. Uporabljeno je bilo nadzorovano strojno učenje. Rezultati eksperimenta so pokazali, da je uporabljen pristop uspešno ločeval med pristnimi in lažnimi pismi.

2. SORODNA DELA

V članku [10] so avtorji uporabili 66 poslovilnih pisem, od katerih je bilo 33 pristnih in 33 lažnih. V raziskavi je prisostvovalo 11 strokovnjakov s področja mentalnega zdravja in 31 psihiatrov pripravnikov. Zbrana mnenja so primerjali z 9 algoritmi strojnega učenja:

- LMT,
- LinSMO,
- Descision,
- JRip,
- NB,
- PART,
- J48,
- Logistic,
- IB3 in
- OneR.

DOI: https://doi.org/10.26493/978-961-7055-26-9.21-25