Nelleke Oostdijk, Hans van Halteren

Rozpoznávání výhrůžných postů na Twitteru

Úvod a motivace

- cílem je detekovat výhrůžné zprávy na Twitteru bez širšího kontextu
- porovnání statistického a manuálního přístupu
- využití: monitorování napadajícího a ohrožujícího chování a možné právní zásahy

Výchozí znalosti

- tweety jsou omezeny na 140 znaků
- výhrůžka = prohlášení o úmyslu způsobit smrt nebo fyzické zranění jedné nebo více osobám, nebo zničit jejich majetek, zabít nebo zranit jejich domácí zvíře
- obtížnost úkolu – výskyt hříček, vtipů, sarkasmu, ironie, frází obsahujících slova spojená s násilím (*je kunt doodvallen = drop dead = jdi se bodnout*)

Výchozí data

- soubor výhrůžných tweetů – cca 5 tisíc postů shromážděných na www.doodsbedreging.nl (stránka zabývající se výhrůžkami na internetu) → testovací (TTS) a vývojový (TDS) soubor
- náhodně vybrané vzorky holadnského Twitteru → testovací (GTS) a vývojový soubor (GDS)

Manuální přístup

- založeno na jazykové intuici autorů (rodilých mluvčích)
- použití unigramů, bigramů a trigramů
- skip gramy = bigramy a trigramy, kde spolu jednotlivé prvky těsně nesousedí
- 3 129 pozitivních n-gramů (indikují výhrůžku) – nejčastěji slovesa označující nějakou akci, např. *doden = zabít, vermoorden = zavraždit*)
- 13 061 negativních n-gramů (odstraňují případy přegenerování), např. *hart aanval = heart attack = srdeční záchvat, aanval = útok* ve sportovním kontextu)
- přidání spellingových variant

Limity manuálního přístupu

- n-gramy jsou limitovány velikostí (max. n = 3)
- délku mezery ve skip gramech nelze definovat přesně
- negativní n-gramy jsou aplikovány nezávisle na pozitivních, které mají rušit
- spellingové varianty jsou uváděny pouze pro izolované tokeny

Strojový přístup

- automatická statistická identifikace n-gramů indikujících výhrůžku
- systematický výběr n-gramů častěji se vyskytujících ve výhrůžných tweetech než v náhodně vybraných
- kvůli výpočetní náročnosti nelze použít skip trigramy
- trénování na TDS a GDS

Statistický přístup

- systém se učí stupně nadužívání n-gramů
- výpočet středních hodnot a směrodatných odchylek častěji se vyskytujících n-gramů na GDS → nadužívání v TDS → průměrná hodnota = stupeň nadužívání n-gramu
- vybráno 337 tisíc n-gramů
  - referující plánované násilí (*zavraždit, útok, bomba, krk*)
  - cíle násilí
  - zvolání (*wollah = I swear = přísahám*)
  - slova odkazující na budoucnost
  - zájmena, spojky, členy

Vyhodnocení

- na testovacích souborech GTS a TTS
- **manuální systém** rozpoznal 84,8 % tweetů z TDS, 84,7 % z TTS a 79,9 % z GTS
  - pozitivně se projevilo přidání negativních n-gramů a spellingových variant
- **strojový systém** rozpoznal 90 % tweetů z TDS, 90,1 % z TTS, ale pouze 55,8 % z GTS
  - kvalita klesá s jiným rozvrstvením dat

Závěr

- oba systémy jsou použitelné
- možnost spojení přístupů a vzájemného doplňování
- pro strojový systém – větší trénovací soubor, filtrování výsledků pomocí manuálně vytvořených vzorců
- pro manuální přístup – nové vzorce na základě tweetů, které byly rozpoznány strojovým systémem a manuálním nebyly