Radboud Repository

Radboud University Nijmegen

PDF hosted at the Radboud Repository of the Radboud University Nijmegen

The following full text is a publisher's version.

For additional information about this publication click this link.
http://hdl.handle.net/2066/106940

Please be advised that this information was generated on 2020-12-01 and may be subject to change.

De digitale speurhond

Het gebruik van internet en de sociale media heeft de afgelopen jaren een enorme vlucht genomen. Het brede publiek heeft vandaag de dag vrijwel ongelimiteerd toegang tot allerlei informatie en berichtgeving. Nog niet eerder was het zo eenvoudig de actualiteit te volgen, daarop te reageren of in te spelen. Ook bij het onderhouden van sociale contacten zijn de moderne media niet meer weg te denken: we houden elkaar doorlopend op de hoogte van wat we aan het doen zijn, hoe we over dingen denken, we maken afspraken, etc. Je zou bijna geneigd zijn te denken dat er aan deze ontwikkeling alleen maar positieve kanten zitten. Toch zijn er ook nadelen. Voor wie er op uit is, bieden de nieuwe media een enorm platform waarop je ongecensureerd je gang kan gaan, en ook nog eens je identiteit kunt verhullen.

Nelleke Oostdijk
CLST, Radboud Universiteit Nijmegen

Internetsurveillance

De politie monitort al enige tijd het internet op dreigingen gericht tegen personen (bijv. kabinetsleden), objecten (bijv. Schiphol), diensten (bijv. openbaar vervoer) en evenementen (bijv. Nationale Herdenking). De internetsurveillant houdt zich bezig met het identificeren van berichten die vanwege de ernst en de waarschijnlijkheid van de (be)dreiging nadere aandacht van de politie behoeven. Eén van de problemen daarbij is het enorme volume aan data. Nederlanders zijn immers massaal actief op o.a. Facebook, LinkedIn, YouTube en Twitter. Er is dan ook behoefte aan een instrumentarium dat de politie kan helpen internet(be)dreigingen te vinden, te verwerken en te valideren.

Dreigtweets

Dreigtweets zijn een typisch voorbeeld van het soort berichten dat de politie graag en liefst zo snel mogelijk op het spoor zou willen komen. Het verzenden van een dreigtweet staat namelijk gelijk aan iemand offline bedreigen en is strafbaar volgens Artikel 285 van het Nederlandse Wetboek van Strafrecht. Hoewel nog onbekend is om welke aantallen het precies gaat, is al wel duidelijk dat het in de honderden tweets per dag loopt. Zo waren er bijvoorbeeld in de aanloop van de onlangs gehouden Tweede Kamerverkiezingen dagelijks enkele tientallen dreigtweets alleen al aan het adres van Geert Wilders.

Dreigtweetdetectie

In opdracht van het Ministerie van Justitie en Veiligheid werd onlangs door onderzoekers van het CLST een (pilot) project uitgevoerd dat tot doel had een methode te ontwikkelen die ingezet zou kunnen worden om dreigtweets te onderscheiden van niet-dreigtweets. Daarnaast zou de methode ondersteuning moeten bieden bij het inschatten van de ernst en waarschijnlijkheid van de dreiging. Het onderzoek richtte zich specifiek op de Nederlandstalige tweets (incl. Nederlandse straattaal en dialect) die verzonden worden in het Nederlandse domein.

Het onderzoek kent een reeks van uitdagingen. Allereerst gaat het om uiterst korte berichten met een maximale lengte van slechts 140 tekens. Bovendien wijkt het taalgebruik in veel gevallen nogal af van wat we gewend zijn vanuit het standaard Nederlands. Zo zien we een soms nogal eigenzinnig gebruik van de orthografie, een enorme variatie in spelling, typische woordkeuzen en formuleringen.
Een extra complicatie is dat in de huidige opzet van elke tweet afzonderlijk, dat wil zeggen zonder enigerlei context, bepaald moet kunnen worden of het een potentiële dreiging betreft of niet.

Hybride benadering

De gebruikte methode combineert machine learning met een linguïstisch gemotiveerde, regelgebaseerde benadering. De twee benaderingen hebben ieder zo hun sterkten en zwakten. In het geval van machine learning blijft de menselijke inspanning tot een minimum beperkt en kan de computer gebruik maken van 'kennis' opgedaan uit de data,

@...: die wilders moet kapot joh kogel door ze kop #verrijking daarom #PVVop1
@...: moet nu een keer klaar zijn met Geert Wilders. / kill hem
Die geert geeft echt een kogel door ze kop nodig.
Wollah als ik oooit in mn leven Wilders tegen kom djoek ik hem gelijk jood
@... wollah jij gaat dood door een moslim die jou nie meer trek anus likker
Wedden als Geert Wilders zo door gaat hij vroeg of laat word vermoord.

Figuur 1. Voorbeelden van dreigingen aan het adres van Geert Wilders

af bek bom doden **dood** echt **ga** gaan **gaat** kanker kapot keer kil **kill** kk kkr kogel kom komt kop **maak** maken man **mes** moeder moord morgen plan school slaan **steek** steken stoer vallen **vermoord** vermoorden we zie

Figuur 2: Woordenwolk van de meest voorkomende woorden in dreigtweets

kennis die wij als mens niet expliciet voorhanden hebben. Daar staat tegenover dat de computer, om met enig succes automatisch te kunnen leren, grote hoeveelheden trainingsdata nodig heeft en die zijn niet altijd voorhanden. De regel gebaseerde benadering is juist niet afhankelijk van trainingsmateriaal. Het opstellen van de regels gebeurt door de linguïst die daarbij put uit zijn/haar kennis en is daardoor nogal arbeidsintensief. Doordat de regels in meer generieke termen patronen beschrijven, worden tevens voorkomens beschreven die nog niet werden geobserveerd.

De machineleermethode is in dit geval een aangepaste vorm van Linguistic Profiling, een methode die eerder door Hans van Halteren werd ontwikkeld en die met succes werd toegepast o.a. voor auteursherkenning [1]. Vooralsnog blijft de performance van de machine learning component achter bij wat op basis van eerdere ervaringen verwacht mag worden. Dit is te verklaren doordat er nog maar weinig trainingsmateriaal beschikbaar is dat voor de huidige taak geschikt is.

Hoewel de regelcomponent verdere uitbreiding behoeft, blijkt hij toch al zeer effectief. De grootste winst t.o.v. machine learning ligt erin dat de mens ook zonder concreet voorbeeldmateriaal goed in staat is om zeer veel mogelijke verwoordingen van dreiging te omschrijven. Het probleem bij deze aanpak is eerder dat er nog een relatief groot aantal tweets ten onrechte als dreiging wordt aangemerkt. Zo passeerden er bijvoorbeeld in verkiezingstijd nogal wat tweets waarin

- advertentie -

melding gemaakt werd van een aanval op een politicus. In de meeste gevallen was er geen sprake van een echte dreiging, maar ging het om metaforisch taalgebruik. Verdere aanscherping van de regelcomponent is dan ook nodig om dergelijke gevallen in de toekomst van detectie uit te sluiten.

Een en ander is geïmplementeerd in een softwaremodule, informeel aangeduid als 'de dreigingsdetector'.

Gebruikerstest

Hoewel er bij de ontwikkelaars al tijdens het pilot project verschillende ideeën rezen over hoe de dreigingsdetector verder uitgebouwd en verbeterd zou kunnen worden, is ervoor gekozen eerst een gebruikerstest te doen op basis van het resultaat van de pilot, alvorens verder te gaan met de ontwikkeling. Een belangrijke overweging hierbij is dat de introductie van een instrument zoals de dreigingsdetector een enorme impact heeft op de huidige werkwijze van de professionals. Geïntegreerd in groter systeem waarin tweets aan de module worden aangeboden heeft een team van de KLPD op Prinsjesdag de module uitgetest. De evaluatie is hiervan is nog niet afgerond, maar de resultaten zijn bemoedigend.

Doorontwikkeling

Behalve dat de gebruikerstest de effectiviteit van de module laat zien en de nodige feedback ontlokt aan de gebruikers ervan, levert de test ook waardevolle inzichten, onder andere v.w.b. de volumes waar we in de praktijk mee te maken hebben. Daarnaast komt er een belangrijke aanvulling op het beschikbare trainingsmateriaal. Voor de verdere doorontwikkeling van de module zijn de plannen momenteel in voorbereiding. Een uitbreiding naar andere media zoals Facebook en LinkedIn ligt daarbij voor de hand, terwijl ook het betrekken van de context van de berichten en auteursprofielen kan bijdragen aan de duiding van potentiële dreigtweets.

[1)] *http://acl.ldc.upenn.edu/acl2004/main/pdf/183_pdf_2-col.pdf*

Monitoring-tool voor de politie Assen

Masterstudent Jimmy Meijer van de Rijksuniversiteit Groningen ontwikkelt voor zijn afstudeeronderzoek een monitoring-tool voor de politie Assen.

"Ik ben in contact gekomen met de politie van Assen nadat mijn fiets in mei 2012 ontvreemd was uit de schuur. De fiets was teruggevonden en stond op het politiebureau, maar twee maanden later had ik mijn fiets nog steeds niet terug. Ik deed via Twitter mijn beklag over de werkwijze van de politie Assen en aan de hand van de beschrijving bij mijn Twitter-profiel ben ik toen uitgenodigd voor een gesprek over social media op het bureau.

De insteek voor de politie was het 24-7 beschikbaar zijn via social media, waarbij mij direct opviel dat Twitter bij de politie Assen voornamelijk gebruikt wordt als uitgaand kanaal. Na enkele brainstormsessies is besloten om Twitter juist als inkomend kanaal te gebruiken door een monitoring-tool te ontwikkelen.

De eerste opzet voor mijn onderzoek was om te kijken naar relevante tweets voor de politie Assen. Daarvoor moeten de tweets dus binnen de regio Assen vallen. Maar hoe dit te bewerkstelligen? Deze vraag bleek iets te specifiek voor mijn scriptie, dat een meer algemeen karakter moet hebben. De onderzoeksvraag is nu dan ook: 'Hoe kunnen we tweets herkennen die moeten leiden tot een actie?'. Zo'n actie kan ontstaan uit een vraag, een klacht of een opmerking. Mijn scriptie is erop gericht een tool te ontwikkelen dat deze tweets automatisch herkent."