Suid-Afrikaanse Tydskrif vir Natuurwetenskap en Tegnologie
ISSN: (Online) 2222-4173, (Print) 0254-3486

Vergelyking van sentimentanalise-tegnieke vir ’n sosialemedia-grootdatastel

Authors:
I. van der Linde[1]
E. Kotzé[1]
G. Dollman[1]

Affiliations:
[1]Department of Computer Science and Informatics, University of the Free State, South Africa

Corresponding author:
I. van der Linde,
ianvdlinde@gmail.com

How to cite this abstract:
Van der Linde, I., Kotzé, E. & Dollman, G., 2016, ‘Vergelyking van sentimentanalise-tegnieke vir ’n sosialemedia-grootdatastel’, *Suid-Afrikaanse Tydskrif vir Natuurwetenskap en Tegnologie* 35(1), a1404. http://dx.doi.org/10.4102/satnt.v35i1.1404

A comparison of sentiment analysis techniques with a social media big data set. The performance and resource impact of de facto sentiment analysis techniques have not been thoroughly investigated in parallel and distributed environments. This research aims to address this gap in literature to aid in the application of these techniques for big data research.

Die enorme hoeveelheid data wat wêreldwyd geskep word, groei vinnig en stel voortdurend nuwe uitdagings aan informatika wat betref die invoering, verwerking en storing van data. Dié nuwe tendens word ‘grootdatastelle’ genoem en het drie hoofkenmerke: volume (die grootte van die data oorskry die vermoë van relasionele databasisse); spoed (data word teen ’n hoë tempo ingevoer); en verskeidenheid (data word ingevoer vanaf verskillende bronne). Die hantering van grootdatastelle vereis dat rekenaars saamwerk in hoëwerkverrigting-rekenaaromgewings wat na gelang van behoefte uitgebrei kan word. Hierdie rekenaaromgewings benodig ook spesiale databasisse wat die data kan stoor en verwerk en oor meer as een rekenaar versprei is.

’n Nuwe databasiskategorie, bekend as NoSQL, is ontwikkel om probleme wat met grootdatastelle gepaardgaan, op te los (Moniruzzaman & Hossain 2013). Hierdie databasisse maak nie van verhoudings gebruik nie, en implementeer ook nie die hele SQL-standaard nie. NoSQL-databasisse word in vier subkategorieë geklassifiseer: sleutelwaardestoorplek, grafiekdatabasisse, dokumentdatabasisse en breëkolomdatabasisse. Elke tipe beskik oor sekere voor- en nadele wat dit vir bepaalde situasies en doeleindes geskik maak. Hierdie studie gaan die uitwerking van ’n NoSQL-databasis op die werkverrigting van eksterne-sentimentanalise-algoritmes meet. Die studie sal gebruik maak van Apache Cassandra en Apache Accumulo, wat beide breëkolomdatabasisse is (Moniruzzaman & Hossain 2013).

Sentimentanalise dui op die proses waarvolgens die sentiment van ’n dokument outomaties deur middel van ’n algoritme bepaal word (Liu 2012). Dit is veral van belang in die sosiale media (as ’n grootdatabron), gegewe die magdom subjektiewe data wat daarin voorkom. Daar is ’n verskeidenheid van sentimentanalise-tegnieke, waarvan hierdie studie vier sal ondersoek, naamlik: ondersteuningsvektormasjiene (*support vector machines*), die leksikongebaseerde analise, die Naïve-Bayes-analise en kunsmatige neurale netwerke. Hierdie tegnieke sal teen mekaar opgeweeg word op grond van hulle teoretiese en empiriese werkverrigting binne ’n hoëwerkverrigting-rekenaaromgewing.

Die teoretiese vergelyking sal gebaseer word op die algoritmiese kompleksiteit volgens groot-O-notasie. Hierdie kompleksiteit sal geïnkorporeer word in verdere berekeninge om die verwagte toename in werkverrigting te bepaal in ’n hoëwerkverrigting-rekenaaromgewing indien meer rekenaars bygevoeg word. Hierdie berekeninge sal gebaseer wees op Amdahl en Gustafon se wette (Vajda 2011).

Die empiriese vergelyking sal data versamel deur middel van eksperimente wat uitgevoer gaan word. Die data-insameling sluit die volgende maatstawwe in: verwerkingsvereistes, geheuevereistes, kraggebruik, netwerk-impak, akkuraatheid, stoorplekvereistes, opleidingstyd (in die geval van masjienleer-algoritmes) en deursette (die hoeveelheid dokumente wat per tydsinterval geklassifiseer word). Eksperimente sal vir elke kombinasie van die vier tegnieke en twee databasisse op twee tot agt rekenaars uitgevoer word. Die databron vir die eksperimente sal boodskappe (*tweets*) wees wat deur Twitter se gratis *streaming*-koppelvlak beskikbaar gestel word.

Note: A selection of conference proceedings: Student Symposium in Science, 29–30 October 2015, University of the Free State, South Africa. Organising committee: Mr Rudi Pretorius and Ms Andrea Lombard (Department of Geography, University of South Africa); Dr Hertzog Bisset (South African Nuclear Energy Corporation (NECSA); Dr Ernie Langner and Prof Jeanet Conradie (Department of Chemistry, University of the Free State).

Die afhanklike veranderlikes vir die studie sal die reeks vooraf genoemde maatstawwe wees, met die keuse van databasis, tegniek, datastel en die hoeveelheid rekenaars as onafhanklike veranderlikes.

Die navorsing behoort 'n aanduiding te gee wat die vereistes en werkverrigting van verskeie sentimentanalise-tegnieke in hoëwerkverrigting-rekenaaromgewings is.

Literatuurverwysings

Liu, B., 2012, Sentiment Analysis and Opinion Mining, *Synthesis Lectures on Human Language Technologies,* 5(1), May. http://dx.doi.org/10.2200/S00416ED1V01Y201204HLT016

Moniruzzaman, A.B.M. & Hossain, S.A., 2013, 'NoSQL Database: New era of databases for big data analytics- classification, characteristics and comparison', *International Journal of Database Theory and Application,* 6(4), August.

Vajda, A., 2011, *Programming many-core chips,* Springer Science + Business Media, United States http://dx.doi.org.10.1007/978-1-4419-9739-5