UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI
INGEGNERIA INFORMATICA,
MODELLISTICA, ELETTRONICA
E SISTEMISTICA
DIMES

Corso di Laurea Magistrale in
Ingegneria Informatica

Tesi di Laurea

Social stream analysis: l'influenza dei bot sulle elezioni presidenziali americane

Relatori
Prof. Domenico Talia
Ing. Fabrizio Marozzo
Ing. Loris Belcastro

Candidato
Marco Domenicano
Matr. 189208

Anno Accademico 2018/2019

A mia nonna

“La scienza è fatta di dati
come una casa è fatta di pietre.
Ma un ammasso di dati non è scienza
più di quanto un mucchio di pietre
sia una vera casa.”
-Henri Poincaré

Sommario

Elenco delle figure

Introduzione

Nel corso della storia i dati sono sempre stati oggetto di studio da parte dell'uomo, il quale, ha cercato di estrarre informazioni capaci di produrre conoscenza, portando numerosi progressi nel campo della scienza, medicina, economia, etc. Negli ultimi anni, la produzione dei dati è aumentata considerevolmente, complici anche la diffusione di social media, tecnologie mobile e macchine con processi automatizzati connesse in rete. Ed è proprio in rete che questi dati confluiscono, come ad esempio l'attività social di milioni di persone, formando un conglomerato di informazioni dal contenuto eterogeneo (audio, video, JSON, e-mail etc.) con velocità di produzione talmente elevata da rendere difficile analizzarli con tecniche di analisi classiche. Questi dati, che prendono il nome di *Big Data,* oltre all'alta velocità di produzione, eterogeneità e dimensione dei dati, presentano altre caratteristiche che ne rendono difficile la memorizzazione, la gestione e l'elaborazione. Nonostante le difficoltà intrinseche richieste per l'uso dei Big Data, vi è un grande interesse da parte di aziende ed enti governativi per lo sviluppo di tecnologie in grado di analizzarli in tempi ragionevoli ed estrarre, da dati grezzi e non strutturati, conoscenza utile che possa portare benefici economici e sociali.

Sono stati sviluppati numerosi framework e sistemi per analizzare grandi volumi di dati immagazzinati nei *data center,* che, però, intercorrendo anche molto tempo dalla data di raccolta a quella di elaborazione, potrebbero risultare inutili in alcuni ambiti dov'è richiesto un feedback in *real time*. Per tali ragioni, in particolare nell'ultimo decennio, la tecnologia ha visto un notevole progresso nello sviluppo di framework e soluzioni tecnologiche per gestire non solo grandi moli di dati, ma anche l'elaborazione in tempo reale, così da portare l'analisi dei dati ad un livello mai raggiunto prima.

Le tecniche di *Big Data Analytics,* sfruttando anche i progressi nello sviluppo di algoritmi di analisi dei dati paralleli e distribuiti, sono in grado di espandersi verso ambiti di studio prima impensabili come l'andamento dei mercati, *fraud detection*

e analisi dei desideri e delle opinioni dei clienti, che consentono di offrire un feedback istantaneo e un supporto alle decisioni in diversi contesti. Ad esempio, l'analisi dei dati raccolti potrebbe consentire di apportare piccole correzioni nel sistema produttivo, ottimizzando così la produttività e riducendo il tempo di intervento in situazioni di errore, con conseguenti benefici economici.

I social media rappresentano una sorgente immensa di dati, contenenti informazioni sui comportamenti, i gusti e le opinioni delle persone. L'ampia diffusione dei social media, inoltre, ha portato anche allo sviluppo di tecnologie di emulazione del comportamento umano, i *bot*, che possono eseguire operazioni in maniera del tutto automatica e sembrare, all'occhio di un osservatore esterno, utenti normali. L'influenza dei bot in tutti i campi dell'informatica è evidente: infatti, esistono numerose sfaccettature di questo strumento in tutti gli ambiti a partire dai chat bot, passando dai web crawler e terminando nei social bot. Quest'ultima categoria di bot, in particolare, simula il comportamento di un utente *trusted*, rendendola difficilmente individuabile da sistemi di contrasto sviluppati per combatterli. L'analisi della loro influenza all'interno dei social è diventata interesse per la sicurezza nazionale, ad esempio per prevenire attacchi informatici da parte di altre potenze al fine influenzare il pensiero dei cittadini.

L'uso elevato dei social media, in particolare, ha raggiunto soglie elevatissime, instaurando connessioni tra persone più intricate, così da creare una "realtà virtuale" specchio della società moderna, degli usi e costumi di ciascuno di noi. L'uso assiduo di questi strumenti varia dall'ambito lavorativo, sociale e politico. Durante le elezioni, eventi politici e referendum, l'uso massivo delle piattaforme, da parte sia dei partiti sia dei cittadini, è tangibile. Ad ogni evento elettorale o dibattito, infatti, le diverse fazioni in campo si danno battaglia per la raccolta di voti e consenso, anche attraverso una robusta attività mediatica sui social media.

Grazie alla possibilità di analizzare i Big Data raccolti durante le campagne elettorali, i social sono divenuti ambito di ricerca per lo sviluppo di algoritmi in grado di predire il risultato delle elezioni o altro avvenimento politico. Il limite di queste analisi è quello di riuscire a predire dei risultati molto tempo dopo della raccolta dei

dati, rendendo difficili l'attuazione di decisioni in grado di rispondere ad una tendenza negativa dell'opinione pubblica e adattare di conseguenze i programmi politici in base a specifiche tendenze e sentimenti che ha l'elettorato.

Questo lavoro di tesi prevede lo sviluppo di una tecnica per l'analisi in streaming dei dati social al fine di stimare i risultati di un evento politico, analizzando anche come i bot possano influire sui risultati ottenuti. Lo studio si basa su dati delle elezioni presidenziali americane del 2016 per il rinnovo del presidente degli Stati Uniti d'America e contestualmente il rinnovo del parlamento e dei suoi 435 membri. Nello specifico, è stato utilizzato un dataset di circa 8 milioni di tweet, raccolti fino a 40 giorni prima dell'evento. Lo stream di dati è stato simulato utilizzano Apache Kafka, mentre la fase di analisi dello stream è stata implementato utilizzando Apache Storm. Il dataset riguardava i cosiddetti *swing state,* ovvero gli stati dove vi è sempre un'incertezza politica.

Nel **primo capitolo** vengono esplicitate le caratteristiche dei big data, le tecniche di analisi e le architetture sviluppate per la gestione di grandi quantità di dati.

Nel **secondo capitolo** sono studiate i framework e le tecnologie sviluppate per l'analisi di dati in tempo reale, evidenziando le caratteristiche e lo scopo di sviluppo.

Nel **terzo capitolo** sono descritte le principali tecniche di *bot detection* e *user polarization* presenti in letteratura.

Nel **quarto capitolo** verrà proposta una metodologia di analisi di dati politici combinando vari framework e ne verranno mostrati i risultati su un dataset elettorale.

Bibliografia

[1] Hadi, H. J., Shnain, A. H., Hadishaheed, S., & Ahmad, A. H. (2014). Big Data and Five V'S Characteristics. In *IRF International Conference*.

[2] "NIST Big Data Interoperability Framework: Volume 1, Definitions" Sept 2015.

[3] Cielen, D., Meysman, A., & Ali, M. (2016). *Introducing data science: big data, machine learning, and more, using Python tools*. Manning Publications Co..

[4] Wu, X., Zhu, X., Wu, G. Q., & Ding, W. (2013). Data mining with big data. *IEEE transactions on knowledge and data engineering*, *26*(1), 97-107.

[5] Sommerville, I. (2011). Software engineering 9th Edition. *ISBN-10*, *137035152*.

[6] "What is stream processing?" Ververica source: https://www.ververica.com/what-is-stream-processing

[7] Kiran, M., Murphy, P., Monga, I., Dugan, J., & Baveja, S. S. (2015, October). Lambda architecture for cost-effective batch and speed big data processing. In *2015 IEEE International Conference on Big Data (Big Data)* (pp. 2785-2792). IEEE.

[8] Singh, D., & Reddy, C. K. (2015). A survey on platforms for big data analytics. *Journal of big data*, *2*(1), 8.

[9] Mittal, A., Jain, V., & Ahuja, T. (2015). Google file system and hadoop distributed file system-an analogy. *International Journal of Innovations & Advancement in Computer Science*, *4*.

[10] Belcastro, L., Marozzo, F., & Talia, D. (2018). Programming models and systems for Big Data analysis. *International Journal of Parallel, Emergent and Distributed Systems*, 1-21.

[11] Aniello, L., Baldoni, R., & Querzoni, L. (2013, June). Adaptive online scheduling in storm. In *Proceedings of the 7th ACM international conference on Distributed event-based systems* (pp. 207-218). ACM.

[12] Perera, S., & Suhothayan, S. (2015, June). Solution patterns for *real time* streaming analytics. In *Proceedings of the 9th ACM International Conference on Distributed Event-Based Systems* (pp. 247-255). ACM.

[13] Iqbal, M. H., & Soomro, T. R. (2015). Big data analysis: Apache storm perspective. *International journal of computer trends and technology*, *19*(1), 9-14.

[14] Carbone, P., Ewen, S., Fóra, G., Haridi, S., Richter, S., & Tzoumas, K. (2017). State management in Apache Flink®: consistent stateful distributed stream processing. *Proceedings of the VLDB Endowment*, *10*(12), 1718-1729.

[15] Olston, C., & Najork, M. (2010). Web crawling. *Foundations and Trends® in Information Retrieval*, *4*(3), 175-246.

[16] Weizenbaum, J. (1966). ELIZA---a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, *9*(1), 36-45.

[17] Metaxas, P. T., & Mustafaraj, E. (2012). Social media and the elections. *Science*, *338*(6106), 472-473.

[18] Ferrara, E., Varol, O., Davis, C., Menczer, F., & Flammini, A. (2016). The rise of social bots. *Communications of the ACM*, *59*(7), 96-104.

[19] Karataş, A., & Şahin, S. (2017). A Review on Social Bot Detection Techniques and Research Directions. In *Proc. Int. Security and Cryptology Conference Turkey* (pp. 156-161).

[20] Douceur, J. R. (2002, March). The sybil attack. In *International workshop on peer-to-peer systems* (pp. 251-260). Springer, Berlin, Heidelberg.

[21] Cao, Q., Sirivianos, M., Yang, X., & Pregueiro, T. (2012, April). Aiding the detection of fake accounts in large scale social online services. In *Proceedings of the 9th USENIX conference on Networked Systems Design and Implementation* (pp. 15-15). USENIX Association.

[22] Danezis, G., & Mittal, P. (2009, February). SybilInfer: Detecting Sybil Nodes using Social Networks. In *NDSS* (pp. 1-15).

[23] Gong, N. Z., Frank, M., & Mittal, P. (2014). Sybilbelief: A semi-supervised learning approach for structure-based sybil detection. *IEEE Transactions on Information Forensics and Security*, *9*(6), 976-987.

[24] Boshmaf, Y., Muslukhov, I., Beznosov, K., & Ripeanu, M. (2013). Design and analysis of a social botnet. *Computer Networks*, *57*(2), 556-578.

[25] Wang, G., Mohanlal, M., Wilson, C., Wang, X., Metzger, M., Zheng, H., & Zhao, B. Y. (2012). Social turing tests: Crowdsourcing sybil detection. *arXiv preprint arXiv:1205.3856*.

[26] Woolley, S. C. (2016). Automating power: Social bot interference in global politics. *First Monday*, *21*(4).

[27] Chu, Z., Gianvecchio, S., Wang, H., & Jajodia, S. (2012). Detecting automation of twitter accounts: Are you a human, bot, or cyborg?. *IEEE Transactions on Dependable and Secure Computing*, *9*(6), 811-824.

[28] Wang, A. H. (2010, June). Detecting spam bots in online social networking sites: a machine learning approach. In *IFIP Annual Conference on Data and Applications Security and Privacy* (pp. 335-342). Springer, Berlin, Heidelberg.

[29] Davis, C. A., Varol, O., Ferrara, E., Flammini, A., & Menczer, F. (2016, April). Botornot: A system to evaluate social bots. In *Proceedings of the 25th International Conference Companion on World Wide Web* (pp. 273-274). International World Wide Web Conferences Steering Committee.

[30] Chavoshi, N., Hamooni, H., & Mueen, A. (2016, December). DeBot: Twitter Bot Detection via Warped Correlation. In *ICDM* (pp. 817-822).

[31] Dickerson, J. P., Kagan, V., & Subrahmanian, V. S. (2014, August). Using sentiment to detect bots on twitter: Are humans more opinionated than bots?. In *Proceedings of the 2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining* (pp. 620-627). IEEE Press.

[32] Howard, P. N., Duffy, A., Freelon, D., Hussain, M. M., Mari, W., & Maziad, M. (2011). Opening closed regimes: what was the role of social media during the Arab Spring?. *Available at SSRN 2595096*.

[33] Gruzd, A., & Roy, J. (2014). Investigating political polarization on Twitter: A Canadian perspective. *Policy & Internet*, *6*(1), 28-45.

[34] Burnap, P., Gibson, R., Sloan, L., Southern, R., & Williams, M. (2016). 140 characters to victory?: Using Twitter to predict the UK 2015 General Election. *Electoral Studies*, *41*, 230-233.

[35] Marozzo, F., & Bessi, A. (2018). Analyzing polarization of social media users and news sites during political campaigns. *Social Network Analysis and Mining*, *8*(1), 1.

[36] Kennedy, C., Blumenthal, M., Clement, S., Clinton, J. D., Durand, C., Franklin, C., ..., & Saad, L. (2017). An evaluation of 2016 election polls in the US. *American Association for Public Opinion Research, www.aapor. org/Education-Resources/Reports/An-Evaluation-of-2016-Election-Polls-in-the-US.aspx# POLLING AND PROBABILISTIC FORECASTING*.

[37] "Our final map has Clinton winning with 352 electoral votes. Compare your picks with ours." David Lauter and Mark Z. Barabak, published on the website of Los Angeles Times: www.latimes.com.

[38] "The most common hashtags tweeted by Russian trolls" Nikhil Sonnad, published on the website Quartz: qz.com