UNIVERSITÀ DELLA CALABRIA

Dipartimento di Ingegneria Informatica, Modellistica, Elettronica e Sistemistica

Corso di Laurea Magistrale in Ingegneria Informatica

Tesi di Laurea

Sviluppo di una metodologia di hashtag recommendation basata su tecniche di embedding

Relatori:
Prof. ***Paolo Trunfio***
Ing. ***Fabrizio Marozzo***
Ing. ***Riccardo Cantini***

Candidato:
Giovanni Bruno
Matricola 189145

Anno Accademico 2018/2019

Indice

Elenco delle figure

Introduzione

Internet ha ridefinito il modo in cui si vive, ha creato nuove convenzioni e abbattuto le precedenti barriere logistiche. Sono miliardi le persone nel mondo che vi accedono regolarmente per motivi che vanno dal lavoro all'intrattenimento, usufruendone dei contenuti presenti e creandone di nuovi. Tra gli usi di internet, particolare interesse riveste quello offerto dai social media, i quali consentono ad utenti di tutto il mondo di condividere le proprie opinioni ed i propri interessi. Tutte queste funzionalità generano una quantità enorme di dati, i quali possono essere analizzati mediante tecniche di data mining in maniera tale da scoprire nuova conoscenza. La quantità di dati resa disponibile è dunque tale da riferirsi ad essa con il termine di big data. I big data sono caratterizzati in primis dal grande volume di informazioni, ma anche da fattori come la velocità con la quale vengono generati e dalla varietà che li contraddistingue. I campi in cui possono essere usati sono molteplici ma le tecniche tradizionali di elaborazione dei dati risultano poco efficienti in tale contesto. I big data hanno incentivato la nascita di soluzioni scalabili di elaborazione dei dati e fornito nuovi contesti applicativi alle tecniche di machine learning.

In questo lavoro di tesi ci si concentra sui dati generati dai social media, ovvero i social big data [40], e sulle tecniche di elaborazione del linguaggio naturale. In particolare, lo studio condotto si focalizza sui dati generati dai siti di microblogging.

Il microblog è una forma di pubblicazione di piccoli contenuti in un servizio di rete sociale, visibili a tutti o soltanto alle persone della stessa comunità. Con la generazione di grandi quantità di post c'è la necessità di un'efficace categorizzazione e ricerca dei dati. Twitter, uno dei più grandi siti di microblogging, permette agli utenti di utilizzare gli hashtag per categorizzare i loro posts ed

è ultimamente emerso come uno dei sistemi di microblog più diffusi. Milioni di utenti attivi producono una quantità massiccia di tweet, mirati alla diffusione di informazioni e all'interazione sociale. Si tratta di post con un vincolo di lunghezza di 280 caratteri che trattano di una vasta gamma di argomenti, tra cui attività politiche, incarichi personali, argomenti sociali emergenti e contenuti promozionali. Gli hashtag associati ai tweet sono dei tag che consistono di una stringa di caratteri preceduta dal simbolo "#". Essi aiutano nella categorizzazione di tutti i post in base al contenuto e al contesto e sono utilizzati per organizzare i tweet, facilitare la ricerca e diffondere argomenti di tendenza creando comunità istantanee con interessi simili. Nonostante la loro efficacia, la maggior parte dei tweet non contengono hashtag, il che ostacola la qualità dei risultati di ricerca [31]. Per questo l'obiettivo di predire o raccomandare gli hashtag ha catturato notevolmente l'attenzione dei ricercatori. Inoltre, lo stile di scrittura informale e il contesto limitato a causa del vincolo nella lunghezza dei caratteri rende difficile l'analisi dei tweet usando i metodi tradizionali di elaborazione del linguaggio naturale.

Il recente successo delle reti neurali in diversi compiti del Natural Language Processing (NLP) ha accelerato la ricerca nel campo dell'analisi dei social media, come la raccomandazione di hashtag per i tweet [38]. Tra gli ultimi approcci di NLP più promettenti si possono citare i meccanismi di attention e i transformers, i quali costituiscono lo stato dell'arte nei campi della traduzione, della comprensione del testo e più in generale nei task sequence-to-sequence.

Tenendo in considerazione tali fattori si propone una metodologia di raccomandazione degli hashtag sperimentando un approccio di traduzione dal testo dei post di microblogging, ovvero i tweet, all'insieme di hashtag corrispondente. Per riuscire in questo compito si è fatto uso di due spazi semantici di embedding separati: quello dei post, o sentence embedding, e quello degli hashtag, o word/hashtag embedding. Uno spazio di embedding è uno spazio n-dimensionale in cui vengono proiettati gli elementi, in questo caso frasi o parole, di cui si vogliono scoprire le relazioni semantiche. Successivamente si sono combinate le informazioni di entrambi gli spazi mappando lo spazio di sentence embedding in quello di hashtag/word embedding tramite una rete neurale feedforward.

L'efficacia della metodologia proposta nella tesi è stata valutata su un dataset di tweet relativi alle elezioni presidenziali del 2016 negli Stati Uniti, ottenendo

risultati molto promettenti. In particolare, sono state utilizzate due misure prestazionali: *Recall* e *Score*, quest'ultima pesa gli hashtag in base alla loro verosimiglianza. È stata inoltre indagata la capacità del modello di scoprire sui tweet politicamente polarizzati l'argomento principale, modellato come la fazione supportata rispetto ad un insieme di hashtag che sono notoriamente a favore di un candidato specifico.

Questo lavoro di tesi si sviluppa su quattro capitoli. Il primo capitolo descrive le sorgenti dei dati come i social network e come essi siano di notevole impatto nella vita di tutti i giorni. Successivamente si analizzano le caratteristiche dei big data e gli usi in cui sono coinvolti. Il capitolo si conclude analizzando le soluzioni disponibili all'analisi dei social big data.

Il secondo capitolo riporta un'analisi delle attuali tecniche riguardanti i task di topic modeling e hashtag recommendation usate nello stato dell'arte. In esso vengono descritte varie tecniche di clustering, classificazione e modelli generativi, con particolare riguardo alle tecniche di elaborazione del linguaggio naturale e alle reti neurali. Infine si descrivono brevemente dei lavori dello stato dell'arte.

Il terzo capitolo espone la metodologia proposta evidenziandone i componenti principali. Si analizzano i modelli di embedding usati come Word2Vec e Google Universal Sentence Encoder, descrivendo nello specifico la rete neurale adottata per ottenere i risultati attesi.

Nel quarto e ultimo capitolo si descrivono le valutazioni effettuate sul sistema di hashtag recommendation proposto. Si evidenziano i risultati ottenuti dalla metodologia con un punteggio medio di *Recall* del 76% e uno *Score* medio del 70%, confermando inoltre la validità del sistema su più dataset. Nel capitolo si presenta un esempio di applicazione del lavoro di tesi che riguarda la valutazione della polarizzazione politica dei tweet.

Bibliografia

[1] Activation functions in neural networks. https://towardsdatascience.com/activation-functions-neural-networks-1cbd9f8d91d6. Accessed: 2020-03-16.

[2] adam. https://machinelearningmastery.com/adam-optimization-algorithm-for-deep-learning/. Accessed: 2020-02-27.

[3] adam-latest-trends-in-deep-learning-optimization. https://towardsdatascience.com/adam-latest-trends-in-deep-learning-optimization-6be9a291375c/. Accessed: 2020-02-27.

[4] Applicazioni e limiti della classificazione di immagini con reti neurali convoluzionali in dispositivi mobili. http://tesi.cab.unipd.it/51999/1/tesi.pdf. Accessed: 2019-10-15.

[5] A beginner's guide to word2vec and neural word embeddings. https://pathmind.com/wiki/word2vec. Accessed: 2019-11-05.

[6] Clustering. https://www.dataskills.it/tecniche-di-clustering/. Accessed: 2019-10-15.

[7] Dbscan. https://www.developersmaggioli.it/blog/clustering-dbscan/. Accessed: 2019-10-15.

[8] Elaborazione del linguaggio naturale. https://lorenzogovoni.com/elaborazione-del-linguaggio-naturale/. Accessed: 2019-10-19.

[9] facebook. https://www.lifewire.com/what-is-facebook-3486391. Accessed: 2020-02-27.

[10] Flickr. https://en.wikipedia.org/wiki/Flickr. Accessed: 2020-02-27.

[11] Il natural language processing e una sua applicazione: il topic modelling. https://finscience.com/it/news/il-natural-language-processing-e-una-sua-applicazione-il-topic-modelling/. Accessed: 2019-10-19.

[12] instagram. https://www.lifewire.com/what-is-instagram-3486316. Accessed: 2020-02-27.

[13] Introduction to word embedding and word2vec. https://towardsdatascience.com/introduction-to-word-embedding-and-word2vec-652d0c2060fa. Accessed: 2019-11-05.

[14] Intuitive guide to latent dirichlet allocation. https://towardsdatascience.com/light-on-math-machine-learning-intuitive-guide-to-latent-dirichlet-allocation-437c81220158. Accessed: 2019-10-14.

[15] linkedin. https://www.businessinsider.com/what-is-linkedin?IR=T. Accessed: 2020-02-27.

[16] Modelli generativi. http://ccnl.psy.unipd.it/research/workshops/modelli-generativi. Accessed: 2019-10-14.

[17] Multi layer perceptron. https://pathmind.com/wiki/multilayer-perceptron/. Accessed: 2020-03-16.

[18] rise-of-social-media. https://ourworldindata.org/rise-of-social-media. Accessed: 2020-02-27.

[19] Tipi di algoritmi per il machine learning. https://www.datawiring.me/tipi-di-algoritmi-per-il-machine-learning/. Accessed: 2019-10-15.

[20] Transformer. https://medium.com/inside-machine-learning/what-is-a-transformer-d07dd1fbec04/. Accessed: 2019-11-05.

[21] twitter. https://www.lifewire.com/what-exactly-is-twitter-2483331. Accessed: 2020-02-27.

[22] Types of classification algorithms in machine learning. https://medium.com/@Mandysidana/machine-learning-types-of-classification-9497bd4f2e14. Accessed: 2019-10-15.

[23] Understanding rnn and lstm. https://towardsdatascience.com/understanding-rnn-and-lstm-f7cdf6dfc14e. Accessed: 2019-10-15.

[24] Vanishing gradient problem. https://towardsdatascience.com/the-vanishing-gradient-problem-69bf08b15484. Accessed: 2020-03-16.

[25] Youtube. https://en.wikipedia.org/wiki/YouTube. Accessed: 2020-02-27.

[26] Loris Belcastro, Riccardo Cantini, Fabrizio Marozzo, Domenico Talia, and Paolo Trunfio. Discovering political polarization on social media: A case study. In *The 15th International Conference on Semantics, Knowledge and Grids*, Guangzhou, China, 2019. To appear.

[27] Loris Belcastro, Riccardo Cantini, Fabrizio Marozzo, Domenico Talia, and Paolo Trunfio. Learning political polarization on social media using neural networks. *IEEE Access*, 8(1):47177–47187, 2020.

[28] Gema Bello-Orgaz, Jason J Jung, and David Camacho. Social big data: Recent achievements and new challenges. *Information Fusion*, 28:45–59, 2016.

[29] Nada Ben-Lhachemi and El Habib Nfaoui. Using tweets embeddings for hashtag recommendation in twitter. *Procedia Computer Science*, 127:7–15, 2018.

[30] Daniel Cer, Yinfei Yang, Sheng yi Kong, Nan Hua, Nicole Lyn Untalan Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Céspedes, Steve Yuan, Chris Tar, Yun hsuan Sung, Brian Strope, and Ray Kurzweil. Universal sentence encoder. In *In submission to: EMNLP demonstration*, Brussels, Belgium, 2018. In submission.

[31] Fréderic Godin, Viktor Slavkovikj, Wesley De Neve, Benjamin Schrauwen, and Rik Van de Walle. Using topic models for twitter hashtag recommendation. In *Proceedings of the 22nd International Conference on World Wide Web*, pages 593–596. ACM, 2013.

[32] Yeyun Gong, Qi Zhang, and Xuanjing Huang. Hashtag recommendation for multimodal microblog posts. *Neurocomputing*, 272:170–177, 2018.

[33] Yuyun Gong and Qi Zhang. Hashtag recommendation using attention-based convolutional neural network. In *IJCAI*, pages 2782–2788, 2016.

[34] Jiajia Huang, Min Peng, and Hua Wang. Topic detection from large scale of microblog stream with high utility pattern clustering. In *Proceedings of the 8th Workshop on Ph. D. Workshop in Information and Knowledge Management*, pages 3–10. ACM, 2015.

[35] Mohit Iyyer, Varun Manjunatha, Jordan Boyd-Graber, and Hal Daumé III. Deep unordered composition rivals syntactic methods for text classification. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1681–1691, 2015.

[36] Zhao Jianqiang, Gui Xiaolin, and Zhang Xuejun. Deep convolution neural networks for twitter sentiment analysis. *IEEE Access*, 6:23253–23260, 2018.

[37] Abhay Kumar, Nishant Jain, Suraj Tripathi, and Chirag Singh. From fully supervised to zero shot settings for twitter hashtag recommendation. *arXiv preprint arXiv:1906.04914*, 2019.

[38] Yang Li, Ting Liu, Jingwen Hu, and Jing Jiang. Topical co-attention networks for hashtag recommendation on microblogs. *Neurocomputing*, 331:356–365, 2019.

[39] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[40] Ekaterina Olshannikova, Thomas Olsson, Jukka Huhtamäki, and Hannu Kärkkäinen. Conceptualizing big social data. *Journal of Big Data*, 4(1):3, 2017.

[41] Ahmed Oussous, Fatima-Zahra Benjelloun, Ayoub Ait Lahcen, and Samir Belfkih. Big data technologies: A survey. *Journal of King Saud University-Computer and Information Sciences*, 30(4):431–448, 2018.

[42] Junbiao Pang, Fei Jia, Chunjie Zhang, Weigang Zhang, Qingming Huang, and Baocai Yin. Unsupervised web topic detection using a ranked

clustering-like pattern across similarity cascades. *IEEE Transactions on Multimedia*, 17(6):843–853, 2015.

[43] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[44] Philip Russom et al. Big data analytics. *TDWI best practices report, fourth quarter*, 19(4):1–34, 2011.

[45] Jieying She and Lei Chen. Tomoha: Topic model-based hashtag recommendation on twitter. In *Proceedings of the 23rd International Conference on World Wide Web*, pages 371–372. ACM, 2014.

[46] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.