Utilizzare l'Intelligenza Artificiale per Combattere le Fake News

Beatrice Portelli[1], **Enrico Santus**[2], **Giuseppe Serra**[1], **Carlo Tasso**[1]
[1] Laboratorio di Intelligenza Artificiale di Udine (AILAB-Udine),
[2] Computer Science and Artificial Intelligence Lab (MIT-CSAIL)
portelli.beatrice@spes.uniud.it, esantus@mit.edu, giuseppe.serra@uniud.it,carlo.tasso@uniud.it

Abstract

Negli ultimi anni si è assistito ad un'esplosione della quantità di notizie commentate e condivise sui *social network*. Quelle con una diffusione più rapida e virale sono spesso le *fake news*, notizie false e non affidabili. Il loro riconoscimento è un compito dal grande impatto politico e sociale e necessita di prendere in considerazione caratteristiche che vanno oltre stile e linguaggio, considerando anche contesto e prove effettive. UNIUD-AILAB e MIT-CSAIL stanno lavorando congiuntamente ad un sistema basato su Reti Neurali che, sfruttando le informazioni sul Web, stima l'affidabilità di un articolo (assegnandogli uno score) e produce le evidenze a favore e contrarie[1].

1 Introduzione

Negli ultimi anni si è assistito ad un'esplosione della quantità di notizie di cronaca commentate e condivise sui *social network*. Sebbene questa pratica abbia degli aspetti positivi, il dibattito che ne deriva è stato contaminato dalla diffusione di notizie non affidabili, alle quali generalmente ci si riferisce con il termine *fake news*.

Dato che questi contenuti creati a fini dolosi hanno un fortissimo impatto politico e sociale nel mondo reale, la comunità del Natural Language Processing (NLP) è stata chiamata in causa al fine di proporre algoritmi per la loro identificazione.

La maggior parte dei lavori presentati fino ad oggi fanno leva sulle peculiarità stilistiche e linguistiche dei testi delle *fake news* (come eccessiva enfasi ed espressioni iperboliche). Tuttavia, con il passare del tempo, le *fake news* tendono a diventare stilisticamente e linguisticamente sempre più simili alle notizie reali, facendo sì che il controllo dei loro contenuti (*fact checking*) resti l'unico approccio affidabile per isolarle.

2 Related Work

Dai primi studi di [Ott *et al.*, 2011] sul rilevamento di opinioni ingannevoli nelle recensioni, l'attenzione della comunità del NLP sulla *deception detection* [Mihalcea e Strapparava, 2009] e sulle *fake news* [Wang, 2017] è cresciuta rapidamente. La gran parte dei lavori iniziali si è concentrata su approcci stilistici [Feng *et al.*, 2012] e linguistici [Pérez-Rosas e Mihalcea, 2015]. Malgrado le buone prestazioni ottenute su *dataset* sintetici, questi metodi falliscono se applicati a dati reali. Nel 2014, [Vlachos e Riedel, 2014] hanno rilasciato un *dataset* sulle *fake news* contenente 211 asserzioni, suggerendo di utilizzarlo per la classificazione di notizie e il rilevamento di similarità con asserzioni già verificate. Nel 2016 [Ferreira e Vlachos, 2016] hanno introdotto un progetto più consistente per il *rumor debunking* (sfatare dicerie): il *dataset* contiene 300 asserzioni dubbie e 2 595 articoli, che possono supportare, refutare o discutere l'asserzione alla quale si riferiscono. Gli autori hanno proposto un classificatore di regressione logistica che utilizza caratteristiche ottenute esaminando il titolo dell'articolo e la sua concordanza con l'asserzione.

Recentemente [Wang, 2017] ha pubblicato "Liar, Liar Pants on Fire", che include 12 836 brevi asserzioni annotate con valore di verità, argomento, contesto, e dati sull'oratore, tra cui il numero di affermazioni vere/false dette in passato dallo stesso. L'autore ha anche proposto un classificatore di *fake news* basato su reti neurali convoluzionali, che fa leva su schemi linguistici superficiali.

Infine, [Thorne *et al.*, 2018] hanno rilasciato FEVER, un *dataset* su larga scala per l'estrazione e la verifica di fatti (*Fact Extraction and VERification*) che consiste di 185 445 asserzioni. Gli autori hanno implementato un approccio a *pipeline*, che recupera informazioni da Wikipedia e raggiunge nel miglior caso un'accuratezza del $50,91\%$.

3 Progetto

L'obiettivo di questo progetto è sviluppare un sistema in grado di predire l'affidabilità di una notizia tenendo conto di qualità sia intrinseche (stilistiche e linguistiche) che estrinseche (contesto, evidenze e credibilità della fonte) dell'articolo. Per raggiungere questo obiettivo, il consorzio sta lavorando ad una soluzione innovativa basata sulle reti neurali, che hanno recentemente ottenuto risultati eccellenti in altri settori di ricerca del NLP (ad esempio in *sentiment classification*, *machine translation*, etc.).

3.1 Sfide Progettuali

Un sistema di questo tipo deve rispondere a diverse sfide. Prima di tutto, come notato da [Vlachos e Riedel, 2014], non può essere un semplice classificatore, dato che avrà bisogno di basarsi su conoscenze del mondo reale. In secondo luogo, non

[1]Progetto finanziato da MISTI's Global Seed Funds program.

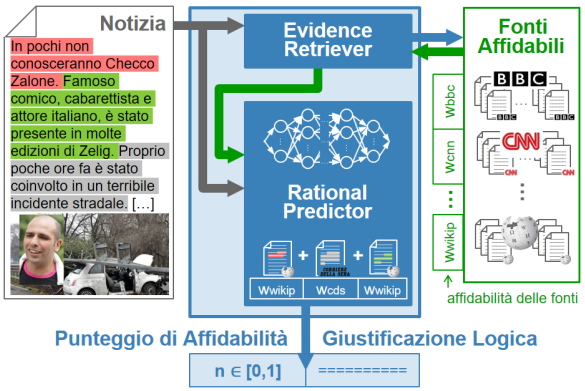

Figura 1: Schema del Sistema Proposto.

può basarsi soltanto sulla similitudine con notizie già controllate, perché queste non sempre sono disponibili o sufficienti. Dovrà quindi ragionare su informazioni recuperate da fonti esterne valutandone l'affidabilità (tenendo a mente che anche fonti considerate "affidabili" possono incorrere in errori).

Inoltre l'affidabilità di una notizia dovrebbe essere misurata con una scala "continua", piuttosto che utilizzando valori binari.

Infine, il sistema dovrebbe restituire una serie di giustificazioni ed evidenze a supporto di ogni predizione, che si basino sia sulla conoscenza del testo (intrinseca) che su quella del mondo (estrinseca). Infatti, come evidenziato in [Lei *et al.*, 2016] l'utilizzo di strategie basate su reti neurali portano a un guadagno in prestazioni, che però è accompagnato a una sempre minore interpretabilità. A tal fine, gli autori hanno proposto un *sentiment analysis* che estrae congiuntamente le predizioni (sentimento positivo e negativo) e le giustificazioni di tale decisione evidenziando un sottoinsieme di parole. Tuttavia, nel nostro scenario, le giustificazioni non possono solamente essere definite come un sottoinsieme del testo originale, ma necessitano dell'inclusione di effettive prove estrinseche, eventualmente ottenute dal Web.

3.2 Architettura

In questo progetto miriamo a creare un'architettura per individuare le *fake news*, che sia al contempo in grado di dare predizioni accurate e di fornire prove intrinseche ed estrinseche a sostegno di tali predizioni. L'obiettivo è di fornire agli utenti la giustificazione logica a supporto della predizione, così che possano leggere la notizia in modo critico.

In particolare, la soluzione proposta (vedi Fig. 1) consisterà di due componenti principali che lavoreranno in sinergia: un modulo dedicato al recupero di fonti esterne di informazioni (*Evidence Retriever*) e un modulo che analizza dinamicamente il testo originale e quello proveniente dall'esterno per restituire un punteggio di affidabilità e identificare simultaneamente la ragione logica per tale predizione (*Rational Predictor*). Modelleremo l'*Evidence Retriever* come un processo decisionale di Markov [Sutton e Barto, 1998], grazie al quale il modello genera ed esegue *query* per raccogliere fonti esterne che sono utili a ottenere predizioni accurate e ben giustificate. Ad ogni passo, dopo aver ottenuto un insieme di fonti esterne, il sistema chiamerà il *Rational Predictor*, che è modellato come un'architettura codificatore–decodificatore [Sutskever *et al.*, 2014] e prenderà in input il testo originale e le fonti recuperate per predire l'affidabilità della notizia e congiuntamente produrre una distribuzione di probabilità sulle parole in input.

Data la natura della soluzione da noi proposta, la strategia di apprendimento sarà modellata come un gioco cooperativo: l'*Evidence Retriever* imparerà come eseguire *query* in grado di raggiungere buone fonti per soddisfare il *Rational Predictor*; il *Rational Predictor* ottimizzerà il punteggio di affidabilità dell'articolo e l'insieme di giustificazioni significative. Questa ottimizzazione può essere formulata in maniera antagonista [Goodfellow *et al.*, 2014].

Riferimenti bibliografici

[Feng *et al.*, 2012] Song Feng, Ritwik Banerjee, e Yejin Choi. Syntactic stylometry for deception detection. In *ACL*, 2012.

[Ferreira e Vlachos, 2016] William Ferreira e Andreas Vlachos. Emergent: a novel data-set for stance classification. In *HLT-NAACL*, 2016.

[Goodfellow *et al.*, 2014] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, e Yoshua Bengio. Generative adversarial networks. *CoRR*, abs/1406.2661, 2014.

[Lei *et al.*, 2016] Tao Lei, Regina Barzilay, e Tommi S. Jaakkola. Rationalizing neural predictions. In *EMNLP*, 2016.

[Mihalcea e Strapparava, 2009] Rada Mihalcea e Carlo Strapparava. The lie detector: Explorations in the automatic recognition of deceptive language. In *ACL/IJCNLP*, 2009.

[Ott *et al.*, 2011] Myle Ott, Yejin Choi, Claire Cardie, e Jeffrey T. Hancock. Finding deceptive opinion spam by any stretch of the imagination. In *ACL*, 2011.

[Pérez-Rosas e Mihalcea, 2015] Verónica Pérez-Rosas e Rada Mihalcea. Experiments in open domain deception detection. In *EMNLP*, 2015.

[Sutskever *et al.*, 2014] Ilya Sutskever, Oriol Vinyals, e Quoc V. Le. Sequence to sequence learning with neural networks. In *NIPS*, 2014.

[Sutton e Barto, 1998] Richard S. Sutton e Andrew G. Barto. Reinforcement learning: An introduction. *IEEE Transactions on Neural Networks*, 16:285–286, 1998.

[Thorne *et al.*, 2018] James Thorne, Andreas Vlachos, Christos Christodoulopoulos, e Arpit Mittal. Fever: a large-scale dataset for fact extraction and verification. In *NAACL-HLT*, 2018.

[Vlachos e Riedel, 2014] Andreas Vlachos e Sebastian Riedel. Fact checking: Task definition and dataset construction. In *LTCSS@ACL*, 2014.

[Wang, 2017] William Yang Wang. "liar, liar pants on fire": A new benchmark dataset for fake news detection. In *ACL*, 2017.