UNIVERSITÀ DEGLI STUDI DI BARI ALDO MORO

Corso di Laurea in Informatica
Elaborato finale di Sistemi di elaborazione per l'automazione d'ufficio:

STATO DELL'ARTE SUI METODI DI IDENTIFICAZIONE DELLE FAKE NEWS

"Se puoi controllare l'opinione delle persone, hai il potere assoluto"

PROFESSORE: Giuseppe Pirlo

STUDENTI: Raffaele Monti, Vincenzo Maria Giulio Martemucci, Pierpaolo Ventrella

OBIETTIVO DEL REPORT

Comprendere tutti gli aspetti del fenomeno legato alle **Fake News**, dalla definizione, a come vengono diffuse, in modo da comprendere le motivazioni che portano alla loro creazione e studiare i metodi che la tecnologia mette a disposizione per identificarle e contrastarne la diffusione.

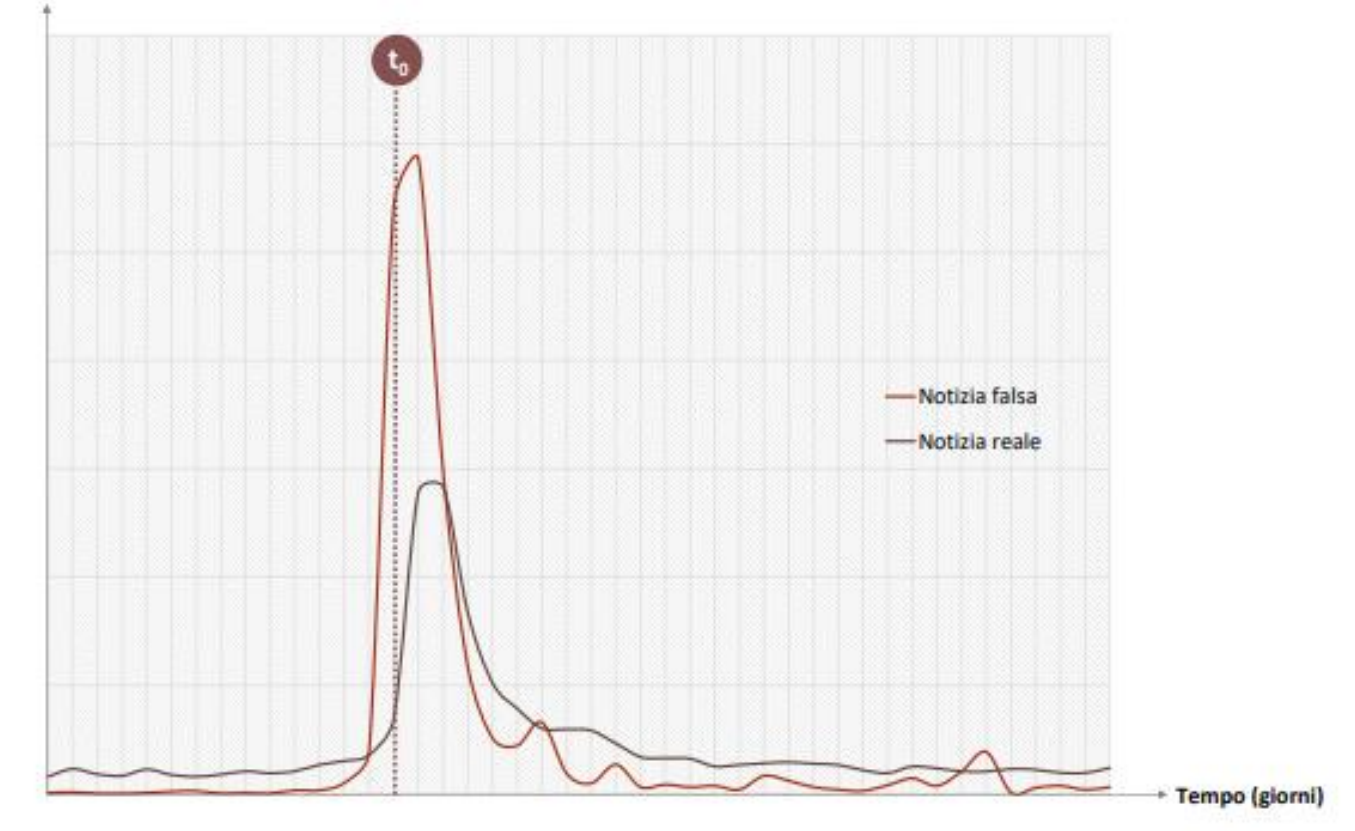
DEFINIZIONE E CARATTERISTICHE

Il termine Fake News, di origine anglosassone, indica articoli redatti con informazioni inventate, ingannevoli o distorte, resi pubblici con il deliberato intento di disinformare o di creare scandalo attraverso i mezzi di informazione.

Caratteristiche principali di una Fake News:

- Stesso pattern
- Titoli accattivanti
- Contenuti studiati ad hoc per suscitare emozioni come indignazione e rabbia.

CICLO DI VITA



Il ciclo di vita di una singola notizia falsa si caratterizza essenzialmente per:

- l'assenza di anticipazioni sui fatti oggetto della notizia falsa;
- una durata sensibilmente inferiore rispetto al ciclo di vita di una notizia reale, con una concentrazione decisamente più accentuata attorno al $t0$, che raggiunge il punto di massimo il giorno successivo al $t0$, per poi scendere velocemente verso valori prossimi allo zero.

TIPOLOGIE DI FAKE NEWS E TECNICHE DI DIFFUSIONE

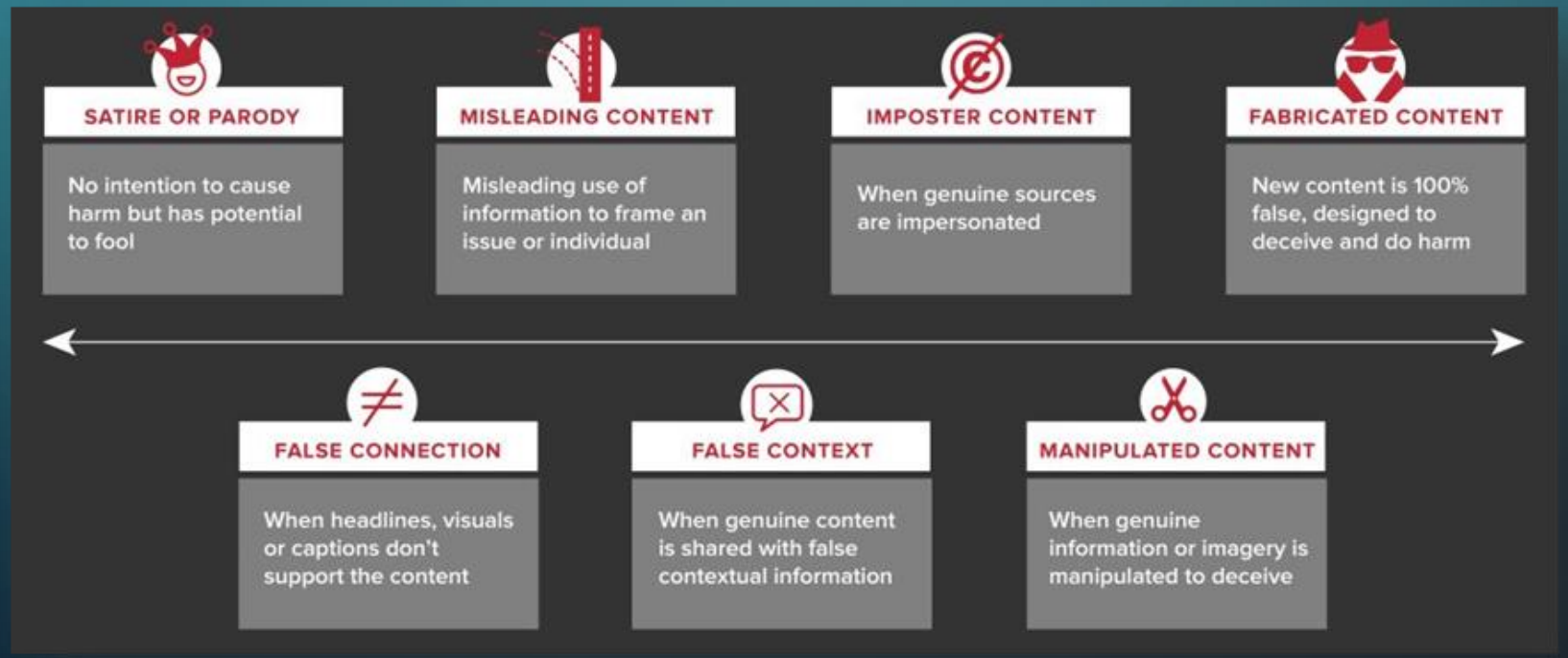
In realtà è il comportamento umano che influisce sulla diffusione di verità o falsità, per questo la comprensione di come le fake news di diffondono è il primo passo per mettere in atto contromisure.

First Draft, la famosa organizzazione no profit che supporta giornalisti, accademici e tecnici nata per combattere la cattiva informazione, suggerisce un metodo di analisi che parte dalla scomposizione dell' ecosistema informativo in tre elementi fondamentali, dai quali far partire l'analisi.

- Tipi di contenuto creato e condiviso
- Motivazioni della creazione del contenuto
- Modalità di diffusione del contenuto

LE DIVERSE TIPOLOGIE DI FAKE NEWS

First Draft ha identificato **sette tipologie** di contenuti informativi problematici che tendono a essere recepiti con difficoltà dagli utenti, considerando la velocità con cui l'approccio all'informazione viene effettuato di questi tempi e quanto affollato sia l'ecosistema informativo sul web e non solo.



LE MOTIVAZIONI DI CHI CREA CONTENUTI DISINFORMATIVI (1)

Le **motivazioni alla base** della “fabbricazione” di Fake news possono essere di vario tipo:

Qualcuno che intende fare profitto, incurante del contenuto dell’articolo falso creato e diffuso

Qualcuno che intende fare satira, per divertimento e quindi con intenti parodici

Giornalisti poco preparati o “costretti” a fornire quante più notizie possibili per via della natura caotica, volatile e repentina del ciclo di notizie odierno

Persone con intenti di parte, con l’intenzione di influenzare l’opinione pubblica e cambiare punti di vista politici

Malintenzionati che intendono guadagnare l’accesso ai dispositivi di chi, cliccando sulla notizia, verrà inconsapevolmente spinto a scaricare dei malware.

LE MOTIVAZIONI DI CHI CREA CONTENUTI DISINFORMATIVI (2)

L'immissione nel sistema informativo di contenuti fake avviene essenzialmente in **tre** passaggi:

- la creazione del messaggio che si vuole trasmettere
- la produzione del contenuto in cui il messaggio viene incorporato e trasformato in un prodotto informativo
- la distribuzione di quest'ultimo.

MATRICE DI DISINFORMAZIONE SECONDO FIRST DRAFT

	SATIRA O PARODIA	COLLEGAMENTO INGANNEVOLE	CONTESTO INGANNEVOLE	CONTENUTO FUORVIANTE	CONTENUTO INGANNATORE	CONTENUTO MANIPOLATO	CONTENUTO FALSO AL 100%
CATTIVO GIORNALISMO		✓	✓	✓			
FARE LA PARODIA	✓				✓		✓
PROVOCAZIONE O PRESA IN GIRO					✓	✓	✓
INTERESSE PARTICOLARE				✓			
FAZIOSITÀ			✓	✓			
PROFITTO		✓			✓		✓
INFLUENZA POLITICA			✓	✓		✓	✓
PROPAGANDA			✓	✓	✓	✓	✓

MODALITÀ E VELOCITÀ DI DIFFUSIONE

Le Fake News viaggiano più velocemente rispetto alla verità. Lo dimostra lo studio realizzato dal gruppo di ricerca del Massachusetts Institute of Technology (MIT) che ha svolto la più vasta analisi su come l'essere umano sparge le notizie, esaminando la piattaforma social Twitter.

I ricercatori hanno analizzato 126mila tweet di circa 3 milioni di utenti pubblicati per più di 4,5 milioni di volte, il tutto in un periodo compreso fra il 2006 e il 2017. Per stabilire se erano veri o falsi, i contenuti sono stati passati al vaglio e confrontati con quelli riportati dalle fonti ufficiali.

La tecnologia sembra essere più d'aiuto a chi intende creare e diffondere le notizie, che a chi vuole difendersi dalle stesse.

Le ultime novità tecnologiche permettono di **generare** e **diffondere** Fake News **in maniera automatica**, attraverso algoritmi di generazione di testo e "bot" che diffondono velocemente ed automaticamente le notizie false.

REPORT DI INFOSFERA

Obiettivo: comprendere quali siano i criteri di scelta delle fonti di informazione degli utenti italiani, quali siano i meccanismi di influenza dei media, in particolare quelli presenti su internet, e la loro efficacia in termini di persuasione.

Lo studio raccoglie i dati provenienti da 1520 questionari somministrati su tutto il territorio nazionale, sulla percezione del sistema mediatico, con particolare attenzione al livello di credibilità, fiducia ed influenza delle fonti di informazione.

RISULTATI DEL REPORT

Il **79,93%** degli italiani ritiene di essere in grado di trovare facilmente le notizie di cui ha bisogno.

L'informazione libera è ritenuta, per **l'87,76%** degli italiani, professionale e, quindi, attendibile.

Per il **93,22%** degli italiani le Fake News hanno impatto nella vita delle persone. Il **65,46%** non riesce a distinguere una Fake News.

Il **78,75%** non è in grado di identificare un sito web di bufale.

L'82,83% non è in grado di identificare la pagina Facebook di un sito di bufale e il **70,28%** non distingue un Fake su Twitter.

L'INCAPACITÀ DI DISTINGUERE IL REALE DAL FALSO

PAGINA FACEBOOK DI UN SITO DI BUFALE

17,17%

È IN GRADO DI IDENTIFICARE IL PROFILO

PROFILO TWITTER FAKE

29,72%

È IN GRADO DI IDENTIFICARE IL PROFILO

SITO WEB DI BUFALE

21,25%

È IN GRADO DI IDENTIFICARE IL PROFILO

FAKE NEWS

34,54%

È IN GRADO DI IDENTIFICARE IL PROFILO

IMPATTO PERCEPITO DEL SISTEMA COMPLESSIVO DI FAKE NEWS

34,80%

SONO CIRCONDATO DA FAKE NEWS

44,54%

IL SISTEMA DEI MEDIA È INVASO DA FAKE NEWS

42,17%

SONO L'EVIDENZA CHE LA RETE È MANIPOLABILE

36,97%

A CAUSA DELLA DISINTERMEDIAZIONE E DELLA MANCANZA DI FILTRI INFORMATIVI AUTOREVOLI IL SISTEMA DEI SOCIAL MEDIA AUMENTA LA DIFFUSIONE DI FAKE NEWS

52,89%

ESISTONO DA SEMPRE MA CON I SOCIAL SONO PIÚ EVIDENTI

32,29%

SONO GENERATE DA TUTTI I MEDIA E NON SOLO DALLA RETE

21,71%

NON MI COLPISCONO

IMPATTO PERCEPITO DELLE FAKE NEWS SUL SISTEMA POLITICO

86,36%

NON HANNO CONDIZIONATO LE ELEZIONI POLITICHE 2018

28,36%

SONO UNA NUOVA FORMA DI PROPAGANDA POLITICA

22,70%

INDEBOLISCONO LA DEMOCRAZIA

CLASSIFICAZIONE DELLE NOTIZIE

Visto il dilagare del fenomeno legato alle Fake News e alla loro diffusione, negli anni sono stati creati diversi **database** di fake news. Uno dei più importanti sia per qualità che per dimensione è quello denominato Fake News Net, creato utilizzando 6 dataset.

Features / Dataset	News Content		Social Context				Dynamic Information
	Linguistic	Visual	User	Post	Second order	Network	
BuzzFeedNews	✓	✗	✗	✗	✗	✗	✗
LIAR	✓	✗	✗	✗	✗	✗	✗
BS Detector	✓	✗	✗	✗	✗	✗	✗
CREDBANK	✓	✗	✓	✓	✗	✓	✗
BuzzFace	✓	✗	✗	✓	✓	✗	✗
FacebookHoax	✓	✗	✓	✓	✓	✗	✗
FakeNewsNet	✓	✓	✓	✓	✓	✓	✓

Partendo da questa mole di dati a disposizione, possiamo definire l'identificazione delle fake news come una funzione F che dato in ingresso un insieme ϵ di notizie predice se esse sono fake o meno.

$$F : \epsilon -> \{0, 1\}$$

$$F(a) = \begin{cases} 1, & \text{se la notizia è una fake news} \\ 0, & \text{se la notizia è autentica} \end{cases}$$

Si individuano 4 classi, che identificano se la news è stata catalogata correttamente.

		Predicted Class	
		No	Yes
Observed Class	No	TN	FP
	Yes	FN	TP

Infine si utilizzano delle metriche per calcolare l'accuratezza e l'affidabilità della classificazione.

$$\mathbf{F_1\ Score} = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$$

$$\mathbf{Precision} = \frac{|TP|}{|TP| + |FP|}$$

$$\mathbf{Recall} = \frac{|TP|}{|TP| + |FN|}$$

$$\mathbf{Accuracy} = \frac{|TP| + |TN|}{|TP| + |TN| + |FP| + |FN|}$$

L'INTELLIGENZA ARTIFICIALE APPLICATA ALLE FAKE NEWS

Riportiamo una serie di studi e soluzioni che hanno ottenuto maggior riscontro nel campo della generazione, rilevazione e identificazione di Fake News:

- Il modello **OpenAI GPT-2**
- L'algoritmo dell'**Università del Michigan**
- La soluzione proposta da **CSAIL**
- Fake News **Detection**

IL MODELLO OPENAI GPT-2

Una nuova branca dell'Intelligenza Artificiale, permette di generare del testo dati un tono ed un argomento.

Tra i vari modelli proposti spicca il modello GPT-2 di OpenAI.

Questo modello è capace di **generare del testo** a partire da una serie di input scritti da un utente umano. Il modello, sulla base di questi input, costruirà un vero e proprio testo, facilmente confondibile con un vero e proprio articolo.

Una Fake News così generata è in grado di imitare la forma e lo stile di notizie autentiche, acquistando credibilità e rendendo molto difficile la distinzione tra notizia vera e notizia falsa.

OpenAI ha sempre rilasciato tutti i suoi prodotti al pubblico, con GPT-2 invece, si è inizialmente scelto di **rilasciare un modello ridotto** proprio per via del potenziale utilizzo del modello per generare Fake News.

Nel 2019 infine, OpenAI ha fatto un passo indietro e in un comunicato ha annunciato che il modello GPT-2 non sarà disponibile al pubblico, in un tentativo di arginare l'utilizzo dello stesso in generazione di Fake News.

UN ESEMPIO DI TESTO GENERATO DA OPENAI GPT-2

SYSTEM PROMPT (HUMAN-WRITTEN)	*Legolas and Gimli advanced on the orcs, raising their weapons with a harrowing war cry.*
MODEL COMPLETION (MACHINE-WRITTEN, FIRST TRY)	The orcs' response was a deafening onslaught of claws, claws, and claws; even Elrond was forced to retreat. "You are in good hands, dwarf," said Gimli, who had been among the first to charge at the orcs; it took only two words before their opponents were reduced to a blood-soaked quagmire, and the dwarf took his first kill of the night. The battle lasted for hours until two of the largest Orcs attempted to overwhelm Aragorn. When they finally stopped, they lay defeated and lifeless for miles and miles. "I take nothing," said Aragorn. "But I give my word, at my peril and mine, that I will never forget this day of horror. None of us will forget. Ever!" "I'll never forget it!" cried Gimli, who had been in the thick of the battle but hadn't taken part in it. One of the wounded orcs he had carried off, he was the only one of the survivors who remained uninjured. "We'll keep the memory of that day of evil, and the war with it, alive as long as we live, my friends!" "Then we'll keep it alive as long as we live," added Legolas. "And we won't forget the first great battle of the night, even if we may have forgotten the final defeat."

L'ALGORITMO DELL'UNIVERSITÀ DEL MICHIGAN

Funzionamento: L'algoritmo utilizza l'elaborazione del linguaggio naturale per cercare modelli specifici e indicazioni linguistiche che segnalano il fatto che un particolare articolo è una notizia falsa. Analizzando punteggiatura, vocaboli, struttura grammaticale e complessità dei testi, il sistema riconosce le spie linguistiche che caratterizzano le notizie false, in modo da fermarle prima che possano generare clic e manipolare l'opinione pubblica.

L'algoritmo è riuscito a individuare le bufale nel 76% dei casi, mentre l'occhio umano normalmente si ferma al 70%.

LA SOLUZIONE DEL CSAIL

CSAIL: Computer Science & Artificial In-telligence Lab

concentrarsi direttamente sulle fonti delle notizie, in maniera tale da poter usare il machine learning per capire se la fonte è accurata o se la stessa ha una parte politica.

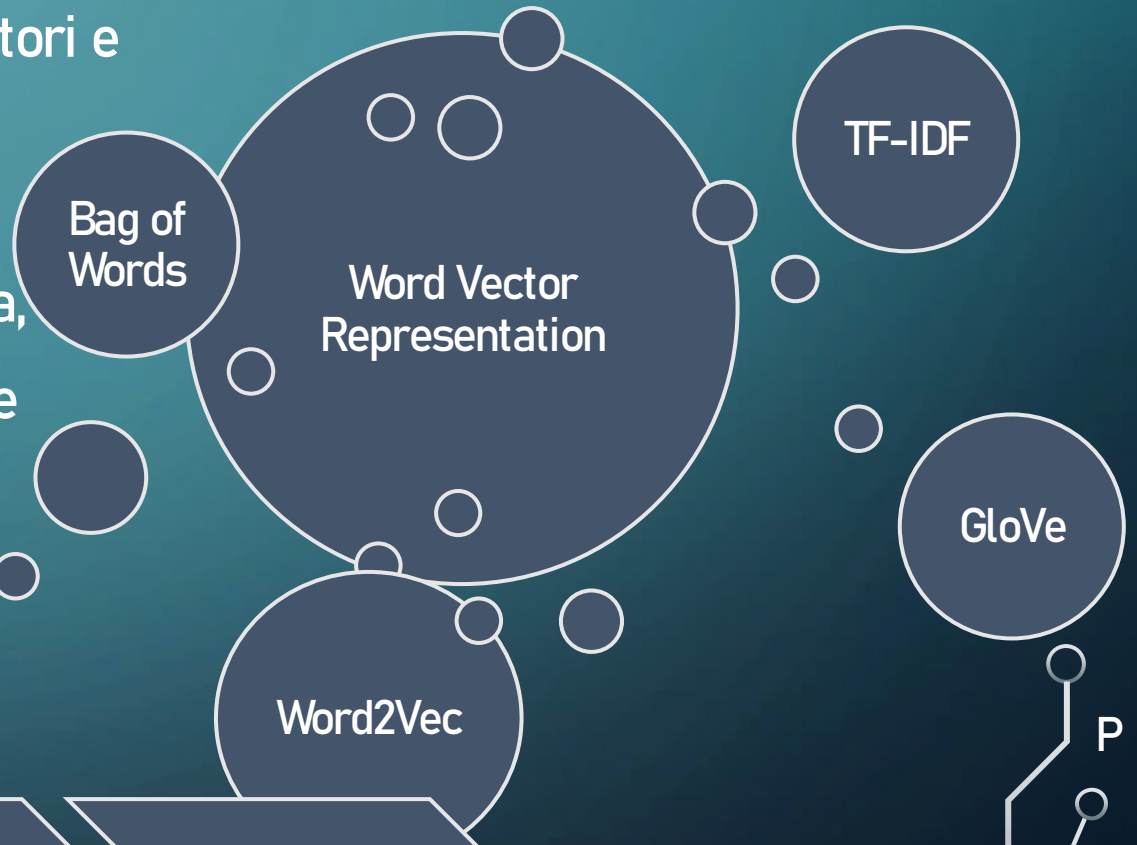
Il sistema ha bisogno solamente di circa 150 articoli per rilevare in modo affidabile se una fonte di notizie può essere attendibile

FAKE NEWS DETECTION

Obiettivo principale del task del Fake News detecting è identificare il linguaggio utilizzato per ingannare i lettori e classificare il testo come reale o fake.

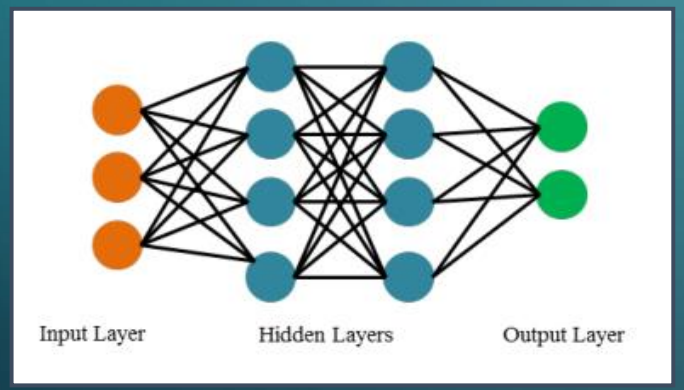
Di grande aiuto è il NLP, che permette di convertire i dati testuali in un formato leggibile da una macchina, utilizzando una rappresentazione delle parole tramite vettore.

Fondamentale diventa il pre-processing del testo, effettuato con diverse tecniche.

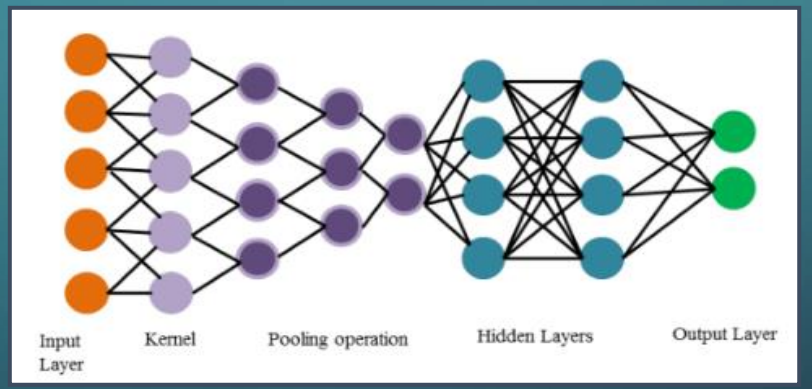


TIPI DI RETI NEURALI

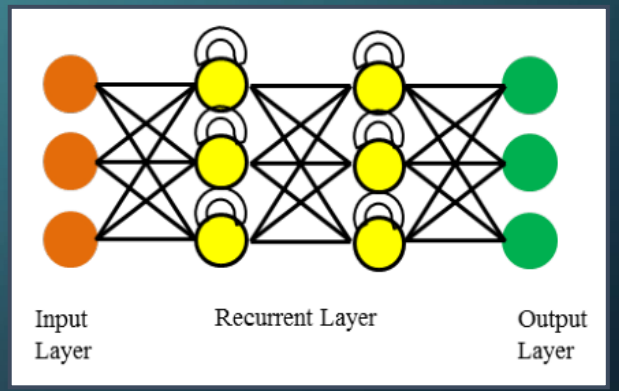
Risulta indispensabile scegliere il giusto tipo di Rete Neurale da applicare al task in questione, analizzandone i benefici che ne possono derivare e considerandone anche l'utilizzo multiplo o parallelo.



Dense Neural Network
- DNN



Convolutional Neural Network
- CNN



Recurrent Neural Network
- RNN

PRINCIPALI TOOL PER L'IDENTIFICAZIONE DI FAKE NEWS

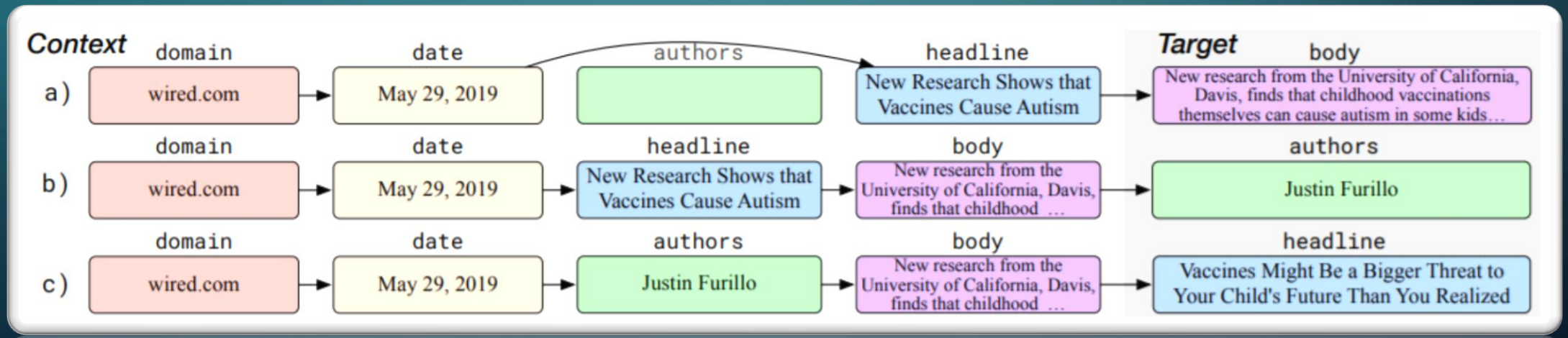
Illustriamo di seguito una **serie di tool** utilizzabili in modo gratuito e non dagli utenti, alcuni dei quali sviluppati in ambienti open source.

- Grover
- Botometer
- Image Self-Consistency tool
- Fakebox

GROVER – ALLEN INSTITUTE FOR AI

Obiettivo: individuare in modo affidabile le cosiddette “Fake News Neurali”

Lo studio alla base è basato sulla rilevazione delle fake news neurali a partire da un modello che le sappia anche generare. Migliora in questo modo l’accuratezza dell’intero Sistema nell’individuare quali notizie sono generate da AI, con una precision di oltre il 92%.



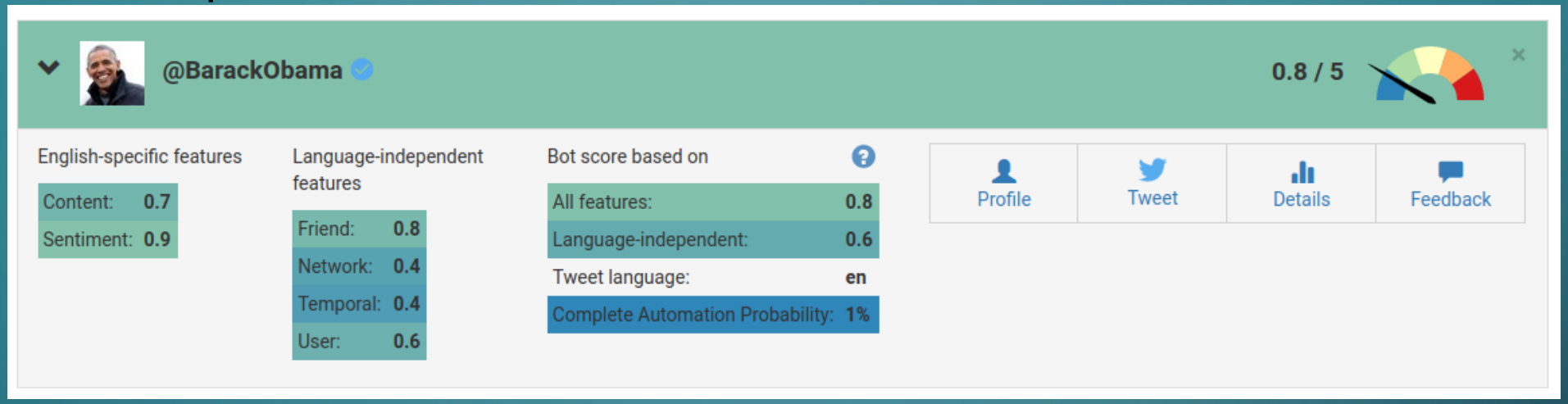
BOTOMETER

Obiettivo: classificare un account sul social network Twitter come bot o umano

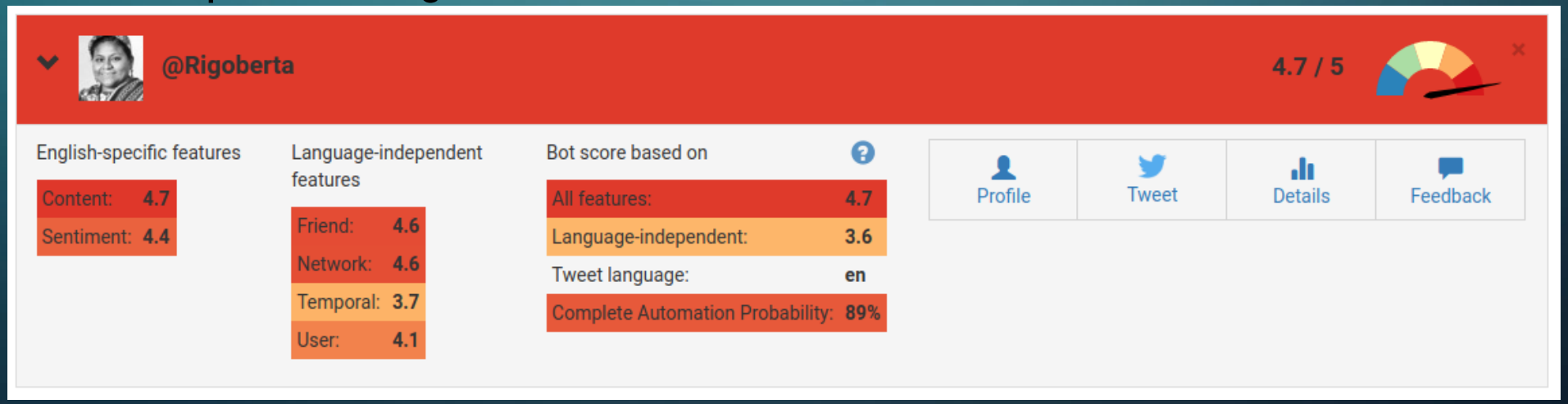
Il tool calcola uno score compreso tra 0 e 5: più alto è lo score, più è alta la probabilità che l'account sia parzialmente o completamente controllato da un software.

Due modalità di utilizzo: tramite sito web o tramite API Python

Analisi di un profilo Twitter autentico



Analisi di un profilo Twitter gestito da un bot



Esempio di codice Python utilizzato per analizzare un account Twitter

```
import botometer

mashape_key = "xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx"
twitter_app_auth = {
    'consumer_key': 'xxxxxxxx',
    'consumer_secret': 'xxxxxxxxxx',
    'access_token': 'xxxxxxxxx',
    'access_token_secret': 'xxxxxxxxxxx',
  }
bom = botometer.Botometer(wait_on_ratelimit=True,
                          mashape_key=mashape_key,
                          **twitter_app_auth)

# Check a single account by screen name
result = bom.check_account('@clayadavis')

# Check a single account by id
result = bom.check_account(1548959833)

# Check a sequence of accounts
accounts = ['@clayadavis', '@onurvarol', '@jabawack']
for screen_name, result in bom.check_accounts_in(accounts):
    # Do stuff with `screen_name` and `result`
```

IMAGE SELF-CONSISTENCY TOOL

I nuovi strumenti di fotoritocco hanno reso molto semplice la creazione di immagini false e rilevare tali manipolazioni è un compito difficile.

I Tool di auto consistenza, utilizzano un algoritmo di apprendimento per il rilevamento di manipolazioni di immagini, che viene addestrato utilizzando un dataset di fotografie reali.

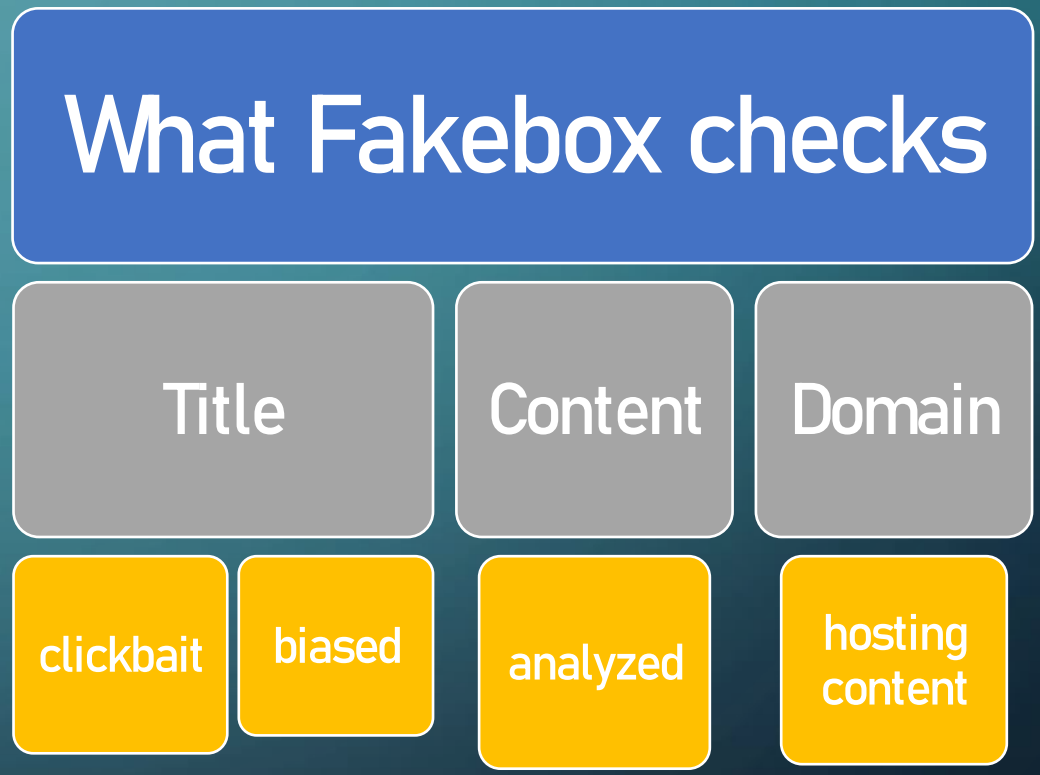
Questo modello di autoconsistenza ha il compito di rilevare e localizzare le giunzioni di immagini, nonostante non abbia mai visto immagini manipolate durante il training.

Un caso clamoroso di manipolazione fotografica avvenne nel 2014, quando il premio Pulitzer Narciso Contreras, manipolò una fotografia scattata in Siria, rimuovendo dalla stessa un'altra fotocamera di un collega che compariva.

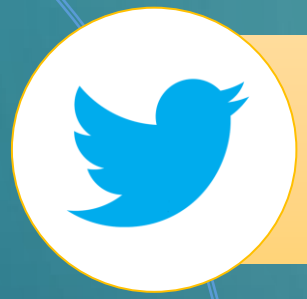
FAKEBOX – MACHINEBOX.IO

Obiettivo: analizzare gli articoli di notizie per valutare se è probabile che siano notizie reali o meno

Osservando una serie di aspetti disponibili di un articolo (titolo, contenuto e URL) e utilizzando modelli di apprendimento automatico integrati e un database curato manualmente, Fakebox può identificare con successo le notizie false.



FAKE NEWS E COVID-19



Concentrarsi su fonti ufficiali quali l'OMS e il CDC nei risultati di ricerca tramite il termine "coronavirus"

Quando le persone cercano informazioni relative al virus si attiverà un pop-up educativo con informazioni attendibili multilingue

Bloccando o limitando gli hashtag utilizzati per diffondere la disinformazione attraverso un popup mirato che in caso di ricerche basate su termini come "coronavirus" riporta alla pagina informativa dell'Organizzazione mondiale della sanità.

NUOVE FUNZIONALITÀ DI WHATSAPP

- Riduzione del numero di volte che i messaggi possono essere inoltrati ai propri contatti con lo scopo di rallentare la diffusione di Fake News.
- È possibile inviare alla piattaforma i messaggi che vengono condivisi riguardanti il Covid-19, per verificare se quanto scritto corrisponde a verità o a bufala facendo riferimento ad altre fonti ufficiali attendibili, fact checker, o attraverso la chatbot dell'International Fact-Checking Network (IFCN) al numero +1 (727) 2912606.
- È possibile salvare il numero di Facta (+39 342 1829843) nei contatti del proprio telefono cellulare e inoltrare a questo i messaggi di testo o vocali, video o immagini dei quali si desidera verificare l'autenticità. Facta, dopo un'attenta analisi, manderà una notifica all'utente che ha inviato la richiesta e, se si tratta di una nuova notizia falsa, la esaminerà e pubblicherà l'analisi sul suo sito web creando un vero e proprio database di notizie false.

CONCLUSIONI

Che il fenomeno delle Fake News sia da contrastare è un'idea condivisibile ma l'impressione è che le soluzioni che si stanno cercando siano dei tappabuchi, che si tratti degli algoritmi o di sanzionare i siti che le pubblicano. Dimenticando che tante di queste news hanno avuto una diffusione enorme grazie a organi di stampa, o esponenti della politica, che le hanno ripubblicate e talora cavalcate a loro vantaggio.

Dunque, più che affidarsi agli algoritmi per valutare l'affidabilità delle news, o accanirsi coi siti che sfruttano il clickbait, sarebbe utile concentrarsi sull'educazione dei cittadini, spronarli a essere più critici, a mantenere un sano scetticismo anche nei confronti delle fonti più autorevoli e dei politici ai quali si sentono più vicini.