

**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

Pierpaolo Ventrella

Vincenzo Maria Giulio Martemucci

# INDICE DEGLI ARGOMENTI

<b>Introduzione .....</b>	<b>pag. 2</b>
 <b><u>Capitolo 1</u></b>	
<b>Definizione di Fake News e ciclo di vita .....</b>	<b>pag. 3</b>
 <b><u>Capitolo 2</u></b>	
<b>Tipologie di Fake news e tecniche di diffusione .....</b>	<b>pag. 5</b>
<ul style="list-style-type: none"><li>• Le diverse tipologie di Fake news</li><li>• Le motivazioni di chi crea contenuti disinformativi</li><li>• Le modalità e la velocità di diffusione delle Fake News</li></ul>	
 <b><u>Capitolo 3</u></b>	
<b>Gli effetti delle Fake News e dati statistici .....</b>	<b>pag. 10</b>
<ul style="list-style-type: none"><li>• Un esempio concreto: il caso Pizzagate</li><li>• Analisi del Report Infosfera</li></ul>	
 <b><u>Capitolo 4</u></b>	
<b>Classificazione delle notizie .....</b>	<b>pag. 13</b>
 <b><u>Capitolo 5</u></b>	
<b>L'intelligenza artificiale applicata alle Fake News .....</b>	<b>pag. 15</b>
<ul style="list-style-type: none"><li>• Un parallelismo con gli Spam detection</li><li>• Uso del Machine Learning per il rilevamento di Fake News<ul style="list-style-type: none"><li>➢ Il modello OpenAI GPT-2</li><li>➢ L'algoritmo dell'Università del Michigan</li><li>➢ La soluzione proposta dal laboratorio CSAIL</li><li>➢ Fake News Detection</li></ul></li></ul>	
 <b><u>Capitolo 6</u></b>	
<b>Principali tool per l'identificazione di Fake News .....</b>	<b>pag. 21</b>
<ul style="list-style-type: none"><li>• Grover</li><li>• Botometer</li><li>• Image self-consistency tool</li><li>• Fakebox</li></ul>	
 <b><u>Capitolo 7</u></b>	
<b>Fake News e Covid-19 .....</b>	<b>pag. 24</b>
 <b>Conclusioni .....</b>	 <b>pag. 26</b>

## INTRODUZIONE

Il 30 ottobre 1938, tramite le onde radio della CBS, la voce dell'annunciatore interrompe il programma musicale: *«Signore e signori, vogliate scusarci per l'interruzione del nostro programma di musica da ballo, ma ci è appena pervenuto uno speciale bollettino della Intercontinental Radio News. Alle 7:40, ora centrale, il professor Farrell dell'Osservatorio di Mount Jennings, Chicago, Illinois, ha rilevato diverse esplosioni di gas incandescente che si sono succedute ad intervalli regolari sul pianeta Marte. Le indagini spettroscopiche hanno stabilito che il gas in questione è idrogeno e si sta muovendo verso la Terra ad enorme velocità.»* Era semplicemente l'inizio di uno sceneggiato radiofonico tratto da "La guerra dei mondi", l'ormai celebre romanzo di H.B. Wells.

La storia creò un panico di massa in quanto molti ascoltatori credettero si trattasse di fatti reali, nonostante fu specificato che la trasmissione non era altro che uno scherzo di Halloween.

Ma fu vero panico? Sembra proprio di no.

Gli ascoltatori anche casuali sono pochissimi. Il programma ha, infatti, uno share del 2%.

Tuttavia, la mattina del 31 ottobre i quotidiani hanno cronache trepidanti: centralini della polizia intasati di telefonate, folle terrorizzate nelle strade, aeroporti chiusi, incidenti stradali, attacchi isterici, suicidi o tentativi di suicidio. Gonfiano il poco, quasi niente che hanno a disposizione fino a dimensioni sensazionalistiche. Letteralmente creano l'immagine del panico.

I quotidiani creano l'evento e ne fanno uno scandalo in un contesto in cui hanno tutto l'interesse a mostrarsi scandalizzati. Il punto focale delle loro storie non è tanto il panico popolare quanto il soggetto che l'ha provocato, cioè la radio. Infatti, in quel periodo la radio si stava affermando sempre più come fonte principale di intrattenimento e informazione, rubando risorse pubblicitarie alla carta stampata. Dunque, quest'ultima coglie l'occasione per screditarla come fornitrice di informazioni affidabili.

Il paradosso, naturalmente, è che per difendere le notizie vere, accurate, affidabili, dai nuovi media elettronici, i print media finirono loro stessi per produrre Fake News.

Si evince quanto le Fake News possano essere determinanti nel veicolare determinati messaggi. Il racconto del panico popolare di allora permane perché è adatto a illustrare i sospetti ricorrenti nei confronti dei nuovi media e della loro capacità di influenzare la vita quotidiana.

Oggi, l'allarme è sollevato da politici, ricercatori e aziende tecnologiche su quello che vedono come una vera e propria epidemia di Fake News.

La diffusione di internet avvenuta nei primi anni 2000 ha aperto nuovi campi a questo fenomeno e la presenza massiccia dei social network nella vita di tutti i giorni, ha fatto da cassa di risonanza.

Se è vero che internet permette di usufruire delle informazioni in maniera facile e quasi istantanea, è anche vero che esso ci espone al pericolo di imbattersi in notizie che non solo sono prive di una fonte autoritaria, ma sono diffuse per scopi sleali, se non illegittimi: ideologici, lucrativi, discriminatori, etc.

Questo problema ha spinto diverse realtà accademiche e istituzionali ad intervenire non solo per analizzare la natura delle notizie che vengono diffuse, ma anche la loro origine, in modo da poter sviluppare mezzi adeguati per contrastarne la diffusione.

Siamo entrati in una nuova fase storica, contraddistinta da una società complessa e instabile, caratterizzata dalla saturazione del tempo di attenzione e dalla frammentazione dei media, che agiscono in maniera pervasiva, determinando il fenomeno del sovraccarico informativo.

In questo nuovo sistema emerge la difficoltà di facoltà critica ed esperienziale nel distinguere il reale dal falso. Pertanto, questo elaborato si pone l'obiettivo di comprendere tutti gli aspetti legati al fenomeno delle Fake News, il che vuol dire definire che tipo di notizie false circolano sulla rete, individuare gli autori, comprenderne le motivazioni e i metodi per contrastare la loro diffusione.

## CAPITOLO 1

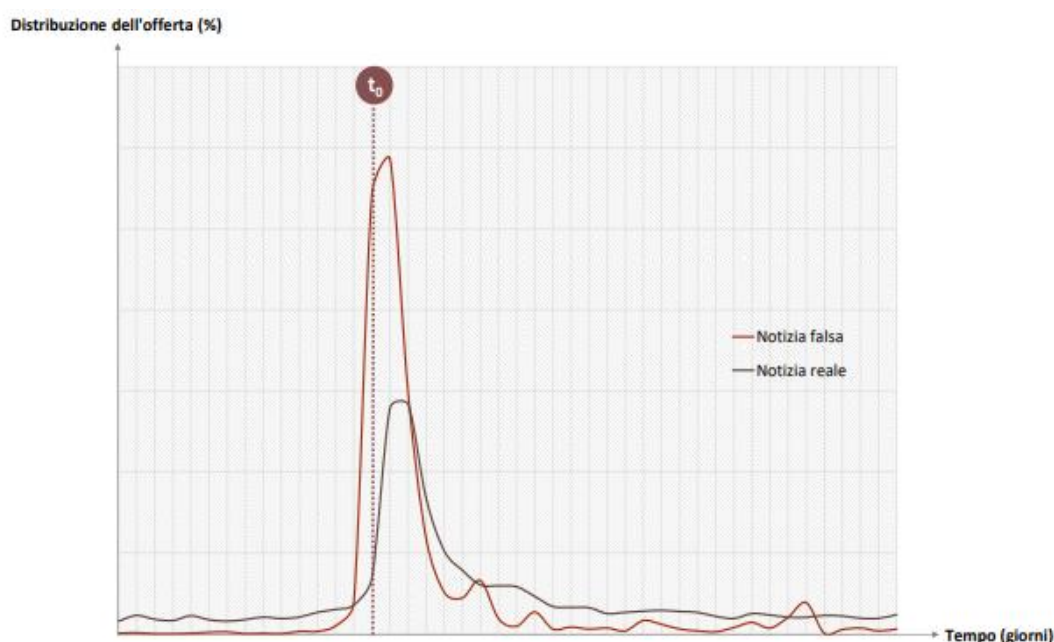
### DEFINIZIONE DI FAKE NEWS E CICLO DI VITA

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. Significativo è in questo senso che i produttori della CBS considerino Fake News tutte quelle storie di cui è possibile dimostrare la falsità ma, nonostante questo, hanno un grande appeal popolare e sono consumate da milioni di persone.

Le Fake News quasi sempre presentano lo stesso pattern, titoli molto accattivanti e contenuti studiati ad hoc per sollecitare emozioni e opinioni (spesso indignazione e rabbia), in modo che le persone le condividano e le facciano diventare virali.

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

- L'assenza di anticipazioni (ossia, contenuti fake diffusi prima del  $t_0$ ) 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  $t_0$ , che raggiunge il punto di massimo il giorno successivo al  $t_0$ , per poi scendere velocemente verso valori prossimi allo zero.



La brevità del ciclo di vita delle notizie false e la concentrazione in pochi giorni della distribuzione dei relativi contenuti sono la spia stessa dell'intento di mettere in atto una strategia di disinformazione, prediligendo la trattazione di tante notizie diverse, evitando di approfondirne i contenuti.

Una volta innescata, la notizia falsa viene immessa e rilanciata nel sistema delle piattaforme online, anche attraverso l'inconsapevole contributo degli utenti, che la condividono e commentano sui social network. Anche la scelta delle tematiche si rivela collimante con l'intento di attivare meccanismi di propagazione virale sulle piattaforme online: in Italia, il 57% della produzione di contenuti fake riguarda argomenti di politica e cronaca, mentre circa il 20% tematiche di carattere scientifico; tutti argomenti che presentano un forte impatto emotivo e che possono essere divisivi, trattati in modo superficiale e impressionistico, mirando a stimolare gli stati d'animo delle persone.

La disinformazione tende ad annidarsi lì dove il sistema dell'informazione fallisce: difficoltà di monetizzazione dei contenuti e contrazione degli investimenti, scarsa preparazione specialistica delle

risorse professionali in determinate materie, esigenze di velocità di aggiornamento dei contenuti informativi (specialmente online) sono atte a compromettere l'adeguatezza dell'offerta informativa sul piano dell'accuratezza, dell'approfondimento e della copertura delle notizie. In termini più ampi, sono alla base della diffusa perdita di reputazione e fiducia accordata dai cittadini al sistema informativo.

Quando le persone sono disinformate, appaiono inclini a difendere le proprie credenze, trascurando le prove concrete. Quando non hanno fiducia nel sistema informativo, il loro atteggiamento di diffidenza le conduce a mostrare resistenza ai fatti, e le correzioni veicolate da fonti ufficiali (seppur scientificamente fondate) possono non riuscire a ridurre le percezioni errate.

È in una situazione di questo tipo che si registra la propensione degli individui ad informarsi affidandosi alla propria rete online di contatti, ad attribuire credibilità ai contenuti e alle fonti che confermano le proprie congetture, a condividere e affermare in prima persona il proprio punto di vista e orientamento ideologico.

Ed è, dunque, su questi atteggiamenti che fanno leva i soggetti che mettono in atto le strategie di disinformazione, contando sull'effetto di viralizzazione che gli stessi possono contribuire a diffondere, unitamente alle caratteristiche tecnologiche e agli algoritmi di personalizzazione delle piattaforme online.

### TIPOLOGIE DI FAKE NEWS E TECNICHE DI DIFFUSIONE

La maggiore probabilità che gli individui diffondano il contenuto falso, non dimostrato o fuorviante postato da altri più della verità, è ciò che guida la diffusione di notizie false consentendogli di proliferare, nonostante la rete favorisca la verità.

Malgrado molte inchieste e rapporti (specie statunitensi) abbiano reso evidente il ruolo dei bot, programmi automatizzati che simulano il comportamento di un utente reale, nella diffusione delle Fake News, altrettanti studi scientifici dimostrano che, in realtà, è il comportamento umano che contribuisce maggiormente alla diffusione di falsità e verità.

Ciò implica che la giusta comprensione di come le Fake News si diffondono sia il primo passo per contenerle e per mettere in atto azioni anche comportamentali efficaci.

First Draft, la famosa organizzazione no profit globale che supporta giornalisti, accademici e tecnici nata per combattere la cattiva informazione e Claire Wardle, la sua direttrice esecutiva, suggeriscono un metodo di analisi interessante. Secondo questo metodo, l'ecosistema informativo può essere analizzato a partire dalla sua scomposizione in tre elementi fondamentali:

- I diversi tipi di contenuti che vengono creati e condivisi
- Le motivazioni di chi crea questi contenuti
- Le modalità di diffusione dei contenuti.

#### Le diverse tipologie di Fake News

Nel settembre del 2016, in occasione delle elezioni presidenziali statunitensi, First Draft ha lanciato un'iniziativa per contrastare le bufale, segnando la prima volta in cui l'espressione Fake News è stata usata nel contesto odierno.

Lo scopo di First Draft era distinguere le notizie verificate dagli elementi di propaganda e sviluppare strumenti perché le prime comparissero in modo prominente nei motori di ricerca e le seconde fossero invece inabissate o eliminate dai selezionatori di contenuti. L'associazione, in particolare, cercava di smontare e sbugiardare le notizie false create ad arte dal febbrile universo di propagandisti e troll della destra americana.

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

Quando si parla di Fake News, in altre parole, si potrebbe parlare di:

- **satira o parodia:** di per sé non ha un valore negativo, anzi molti sistemi giuridici tra cui quello italiano le riservano una garanzia costituzionale come forma d'arte; non di rado succede però che venga manipolata a piacimento, specie per nuocere avversari e contendenti politici;
- **contenuti fuorvianti:** si tratta di informazioni travisate con l'obiettivo di mettere in cattiva luce qualcuno o qualcosa;
- **contenuti falsi:** sono notizie false al 100%, costruite ad hoc per supportare o al contrario recare danno a un'idea, un personaggio, un movimento, una posizione (sono a rigore le vere Fake News);
- **contenuti ingannevoli:** anche in questo caso si tratta di notizie false, attribuite però a fonti realmente esistenti e in genere molto credibili;
- **contenuti manipolati:** sono notizie che, anche se hanno una base di verità, vengono manipolate ad hoc;
- **contesto ingannevole:** non solo il fatto in sé, spesso anche la scelta di una serie di elementi contestuali può risultare faziosa, indurre in errore il lettore e contribuire quindi a creare una Fake News;
- **collegamento ingannevole:** altrettanto ingannevoli possono risultare anche paratesti come le immagini, il richiamo ad altri articoli, etc.

Inoltre, al giorno d'oggi le breaking news o notizie dell'ultimo minuto ed esclusive si stanno rivelando uno dei modi più efficaci per veicolare disinformazione. La logica dello scoop attrae molti giornalisti, ma spesso

è foriera di problematiche piuttosto rilevanti: l'esigenza di tempestività contrasta con la necessaria verifica della veridicità. Anche il click-baiting è una tecnica giornalistica che ben si presta a tattiche di disinformazione poiché prevede l'utilizzo di titoli sensazionalistici, a volte uniti ad immagini in evidenza create apposta per colpire e catturare l'attenzione o diffondere atteggiamenti allarmistici.

### **Le motivazioni di chi crea contenuti disinformativi**

La creazione delle Fake News può avvenire da parte di persone che agiscono singolarmente o in gruppo con lo scopo di rendere le loro "notizie" virali e pervasive o tramite l'utilizzo di bot creati ad hoc.

Tuttavia, nella maggior parte dei casi, dopo la diffusione iniziale avvenuta tramite bot, saranno gli utenti reali a contribuire alla diffusione delle informazioni false, questo in molti casi rende i post più credibili agli occhi di altri utenti reali.

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.

Questi autori, inoltre, mettono in atto vere e proprie strategie per disinformare, essi infatti cercano in molti casi di diffondere notizie false ricoprendole di un costrutto logico coerente e plausibile, spesso distorcendo, per proprio tornaconto, fatti realmente accaduti ma narrati in maniera non corretta, interpretata a proprio favore per orientare (o meglio, disorientare) il lettore.

Lo scopo è quello di strumentalizzare un evento a proprio favore, mescolare quindi verità e menzogna, per depistare, rendere più difficile capire quali fonti siano vere e quali no.

L'immissione nel sistema informativo di contenuti fake avviene essenzialmente in tre passaggi, ossia la creazione del messaggio che si vuole trasmettere, la produzione del contenuto in cui il messaggio viene incorporato e trasformato in un prodotto informativo, e, quindi, la distribuzione di quest'ultimo (*vedi figura a pag. 7*).

Nella **fase di creazione**, viene elaborato il messaggio da veicolare mediante il contenuto fake; questo assume caratteristiche diverse in ragione dell'obiettivo degli ideatori e a seconda del target cui è destinata la strategia di disinformazione. In generale, il messaggio viene ideato in modo da attivare l'audience cui è rivolto, coinvolgendola anche nella diffusione ulteriore del contenuto.

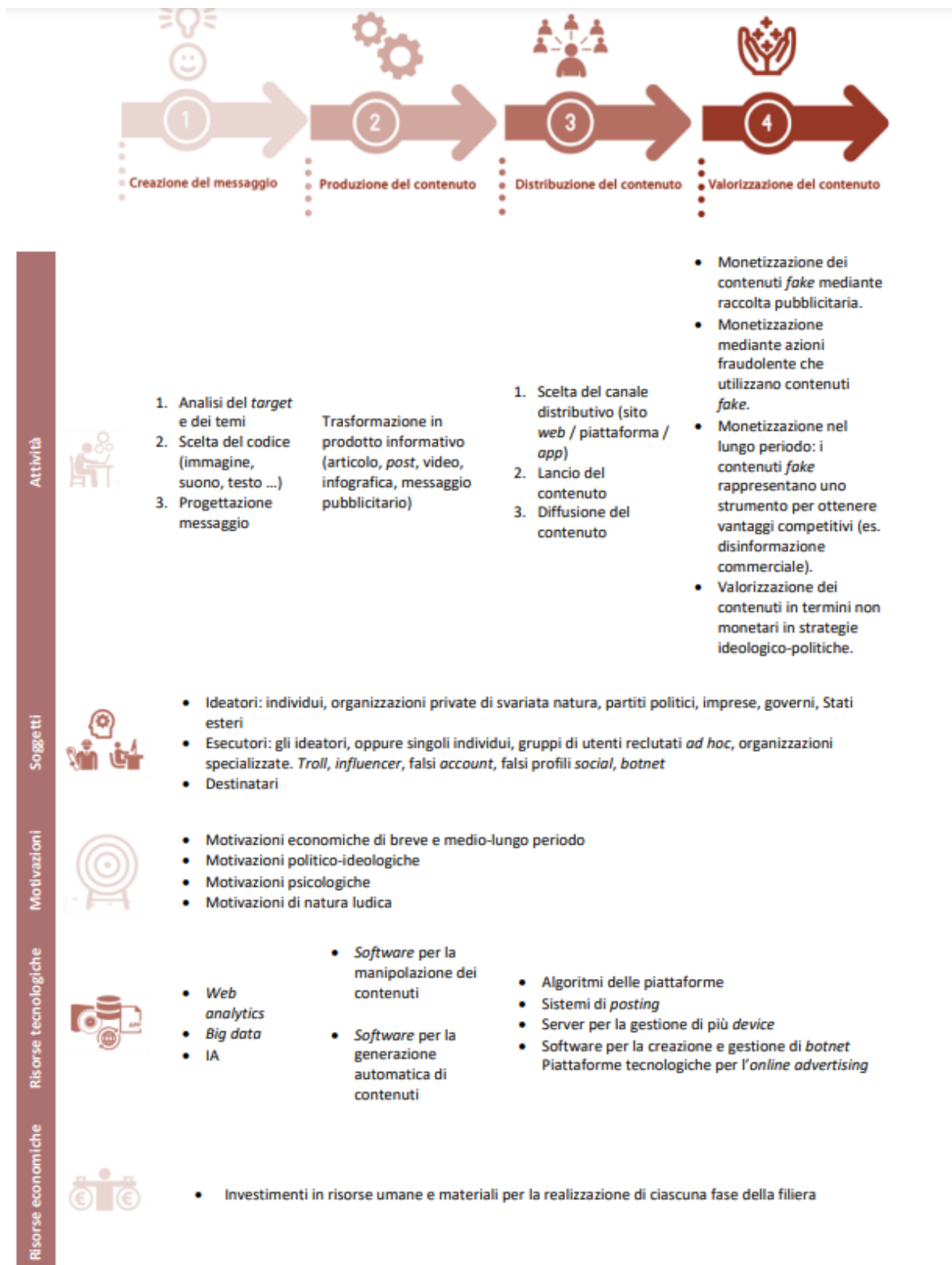
Nella **fase di produzione** del contenuto, il messaggio viene trasformato in un prodotto informativo, che può assumere la forma di un testo (ad esempio, un post o un articolo), di un'immagine, di un video, oppure una combinazione di questi.

Infine, nella **fase di distribuzione**, il contenuto fake viene pubblicato online e reso quindi disponibile. La distribuzione si concretizza attraverso uno o più canali, tipicamente un sito web o una piattaforma online, e consente di collocare il contenuto fake nel contesto mediatico prescelto, che svolge un ruolo importante soprattutto nel conferire attendibilità al messaggio.

Alla realizzazione di contenuti fake partecipano generalmente diversi soggetti. Si distinguono in particolare gli ideatori del contenuto o di un'intera campagna di disinformazione (singoli individui; imprese editoriali e non; organizzazioni con finalità culturali, ideologiche, politiche, criminali; servizi di intelligence; governi; Stati), e gli esecutori delle diverse attività lungo la filiera. Gli esecutori sono coloro che contribuiscono direttamente alla creazione e produzione del contenuto fake, e talvolta coincidono con gli ideatori dell'iniziativa. Possono essere singoli individui, gruppi di utenti reclutati ad hoc, organizzazioni vere e proprie specializzate nella progettazione e implementazione di campagne di disinformazione.

Molto spesso, inoltre, i soggetti che perseguono strategie di disinformazione possono agire con l'ausilio di meccanismi automatici come i bot, che consentono la pubblicazione e distribuzione dei contenuti fake attraverso una molteplicità di account falsi o falsi profili social. Alla divulgazione dei contenuti fake possono

concorrere altresì gli stessi destinatari dei contenuti, laddove, anche inconsapevolmente, si trovano a rilanciarli favorendone la diffusione.



Ci troviamo di fronte ad una sorta di vera e propria filiera produttiva, in cui possiamo identificare nella prima fase, una sorta di operazione di marketing, di progettazione preliminare vincolata ad un obiettivo finale.



Per comprendere al meglio le motivazioni alla base della creazione dei contenuti disinformativi, sempre First Draft ha elaborato uno schema che incrocia 7 modi di fare disinformazione con 8 possibili motivazioni

- Cattivo giornalismo
- Semplice parodia
- Provocazione o presa in giro
- Interesse particolare
- Faziosità
- Profitto
- Influenza politica
- Propaganda

L'intento della matrice è piuttosto evidente: analizzare e comprendere le motivazioni alla base dei vari contenuti può aiutare a combattere la diffusione delle Fake News a cominciare proprio dalla corretta individuazione delle aree di indagine per tipologia di contenuti fuorvianti prescelti.

FIRSTDRAFT		MISINFORMATION MATRIX					
	 SATIRE OR PARODY	 FALSE CONNECTION	 MISLEADING CONTENT	 FALSE CONTEXT	 IMPOSTER CONTENT	 MANIPULATED CONTENT	 FABRICATED CONTENT
POOR JOURNALISM		✓	✓	✓			
TO PARODY	✓				✓		✓
TO PROVOKE OR TO 'PUNK'					✓	✓	✓
PASSION				✓			
PARTISANSHIP			✓	✓			
PROFIT		✓			✓		✓
POLITICAL INFLUENCE			✓	✓		✓	✓
PROPAGANDA			✓	✓	✓	✓	✓

### Le modalità e la velocità di diffusione delle Fake News

L'analisi delle modalità attraverso le quali vengono distribuite le Fake News, può essere utile per comprendere la portata "virale" dei contenuti ed esaminarne i punti di propagazione. Capire come si diffondono le notizie false è infatti un primo passo verso la relativa riduzione.

La quantità di dati e di informazioni prodotta dall'umanità digitale ha raggiunto misure ipersensibili. Una pioggia di notizie in cui si incrociano e si confondono verità e falsità rendendo sempre più complicato verificarne la veridicità e l'autorevolezza.

Nel 2013, il World Economic Forum ha inserito la disinformazione digitale nella lista dei "rischi globali": capace di avere risvolti politici, geopolitici e, perfino, terroristici.

È ormai assodato che 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.

Rispetto alle informazioni veritiere, le Fake News hanno viaggiato più rapidamente, in maniera più radicata all'interno delle diverse categorie di pubblico, e coprendo una maggiore distanza spaziale. Tale fenomeno è accaduto in tutti i campi del sapere anche se, in vetta alla classifica per numero di tweet falsi, vi sono le notizie che riguardano la politica.

I ricercatori si sono chiesti il perché ed hanno messo a fuoco alcuni punti chiave alla base di questo comportamento analizzando 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. In seguito, per stabilire se erano veri o falsi, i contenuti sono stati passati al vaglio e confrontati con quelli riportati dalle fonti ufficiali.

In tutto ciò, la tecnologia sembra, quantomeno in apparenza, essere più d'aiuto a chi intende creare e diffondere le notizie, che a chi vuole difendersi dalle stesse.

La creazione di testi, il copia incolla, la condivisione molto semplice di contenuti online, ha aiutato moltissimo gli autori e i diffusori di queste notizie.

In quasi tutti gli articoli poi, è chiaro un intento provocatorio di chi li diffonde: si vuole causare una forte risposta emozionale in chi li legge, facendo in modo che il lettore stesso, fortemente indignato, diventi poi un diffusore a sua volta della notizia.

Infine, 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.

Il problema non è tanto che siamo attratti da queste storie, ma che divulgandole possiamo fare molti danni, alcuni esempi, il legame vaccini-autismo, il caso Stamina e il più recente riguardante le bufale legate al Covid-19. La diffusione di Fake News sul virus, sulla sua pericolosità o su eventuali rimedi curativi, ha complicato notevolmente la gestione dell'epidemia tanto che l'Organizzazione mondiale della sanità ha chiamato tale fenomeno **Infodemia** indicando l'abbondanza di informazioni, alcune accurate e altre no, che rendono difficile per le persone trovare fonti affidabili quando ne hanno bisogno.

Un caos informativo in cui i fatti accertati e verificati si mescolano alle voci prive di fondamento o ai dati provvisori suscettibili di revisioni anche sostanziali. In un marasma del genere è sempre più difficile orientarsi, perché non si capisce bene a quale delle tante informazioni contraddittorie si debba credere. Il rischio, quindi, è quello di non agire nella maniera più razionale proprio a causa di questa disinformazione diffusa e veicolata dall'essere umano.

### GLI EFFETTI DELLE FAKE NEWS E DATI STATISTICI

#### Un esempio concreto: il caso Pizzagate

Il 4 dicembre 2016, circa a metà giornata, avvenne una sparatoria in una pizzeria nella zona nord-ovest di Washington DC, negli Stati Uniti. La pizzeria si trova in una zona commerciale molto tranquilla di Washington ed in quel momento era frequentata da diverse famiglie. Un uomo, armato di fucile, entrò nella pizzeria ed iniziò a sparare. Fortunatamente, nessuno venne ferito e il sospettato fu arrestato, ma il motivo di questo crimine e le circostanze che lo hanno scatenato furono scioccanti.

La pizzeria, chiamata Comet Ping Pong, fu coinvolta in una strana situazione a causa di un evento avvenuto circa un mese prima. Infatti, ci fu un'elevata diffusione di tweet falsi in rete, che affermavano che la pizzeria era un covo di pedofili dove avvenivano abusi sui minori e che le persone coinvolte, erano vicine alla candidata Presidente democratica Hillary Clinton.

Da quel momento, i proprietari e i dipendenti della pizzeria ricevettero minacce da parte di attivisti di estrema destra che credevano che le notizie fossero vere.

Secondo il Washington Post, a far traboccare il vaso fu l'annuncio del 28 ottobre sulla ripresa delle indagini da parte dell'FBI, sulla questione dell'uso della posta elettronica privata da parte di Hillary Clinton durante il suo mandato come segretaria di stato. Due giorni dopo, un'enorme quantità di tweet riportavano la scoperta di questa banda di pedofili, grazie ad un'e-mail della Clinton.

Questi post vennero diffusi su siti anonimi e tramite social media, assieme a molti video offensivi diretti alla signora Clinton pubblicati su siti di estrema destra. Tuttavia, due giorni prima delle elezioni, l'FBI annunciò che non avrebbe perseguito Hillary Clinton.

I siti anonimi, dunque, focalizzarono la loro attenzione sulla pizzeria, menzionata in diverse e-mail di John Podesta, capo della campagna di Clinton, diffuse sul noto sito WikiLeaks.

Il giorno prima della votazione per le elezioni presidenziali, apparve l'hashtag "#pizzagate".

Anche dopo la sconfitta della Clinton, i tweet non si placarono e continuarono invece ad espandersi.

In seguito, la Central Intelligence Agency (CIA) riferì di aver stabilito che dietro questi attacchi informatici all'e-mail di funzionari del Partito Democratico, come il signor Podesta, vi era un intervento della Russia volto a garantire la vittoria delle elezioni al candidato presidenziale Trump, attuale Presidente degli Stati Uniti.

Man mano che cresceva il numero di persone che credevano nella cospirazione del "pizzagate" e aumentavano le minacce rivolte alla pizzeria, anche altri negozi del quartiere, mai menzionati nei tweet falsi, vennero coinvolti. I dipendenti della pizzeria e delle aziende circostanti affermarono di essere spaventati da un crescente numero di persone che, credendo alle notizie false, frequentavano la zona mostrando un comportamento ostile.

Sebbene i social media abbiano successivamente vietato i post relativi a pizzagate, bloccando i relativi hashtag, le minacce non si fermarono, culminando con la sparatoria di dicembre, in cui un uomo di 28 anni della Carolina del Nord, si presentò al ristorante con un fucile.

Un'intervista del New York Times all'indagato, dopo la sua cattura, descriveva l'uomo come gentile ed educato che intendeva solamente salvare i bambini intrappolati nella pizzeria.

Sebbene nessuno sia rimasto ferito, una vera sparatoria causata da Fake News è un fatto gravissimo e pericoloso per ogni società moderna.

#### Analisi del Report Infosfera

Il Report di Infosfera, svolto dai docenti Umberto Costantini ed Eugenio Lorio dell'università Suor Orsola Benincasa di Napoli, si pone l'obiettivo di 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.

L'analisi parte con un presupposto fondamentale, ovvero la continua diffusione della tecnologia e la presenza di dispositivi digitali connessi al web. Ciò ha portato il singolo individuo a rivalutare completamente la dimensione gerarchica delle fonti di informazione classiche, facendo sì che esse perdessero di credibilità e attendibilità in favore di nuove fonti provenienti dalla rete.

L'infinita libertà di scelta dell'informazione e la presunta possibilità di poter attingere alla conoscenza e alle "vere informazioni" in completa autonomia e senza intermediari, porta gli utenti a costruire palinsesti personali di informazione e conoscenza conformi alle loro attitudini, ai loro gusti, ai loro tempi, alle loro convinzioni. Il 79,93% degli italiani ritiene di essere in grado di trovare facilmente le notizie di cui ha bisogno e tende a fare un largo uso di free media piuttosto che di media a pagamento.

L'alta disponibilità di 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. Inoltre, nonostante circa l'82% degli italiani non riesca a riconoscere facilmente le Fake News, per il 79,93% ognuno è in grado di trovare facilmente le informazioni di cui ha bisogno, andando a confermare ulteriormente il poco peso che viene attribuito al fenomeno.

Il 65,46% non riesce a distinguere una Fake News. Le percentuali crescono quando si tratta di identificare un sito web di bufale, il 78,75% non è in grado di farlo. 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.

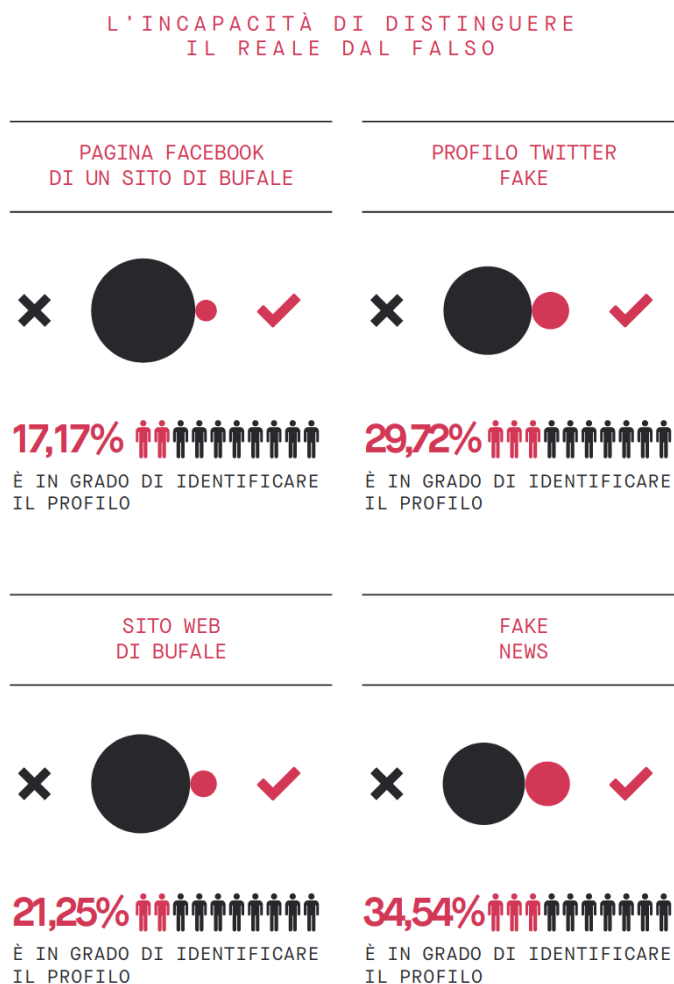
È facile desumere che, ad esempio, l'architettura di Facebook consente ad una notizia che arriva per prima nella nostra timeline, di conquistare più facilmente la nostra attenzione. Di conseguenza il suo valore economico aumenta. Non è dunque la notizia ad essere rilevante bensì le caratteristiche strutturali del mezzo che la veicola. Gli italiani valutano i contenuti con maggiore attenzione quando questi rientrano nella sfera dei loro interessi. Quando ciò non accade, allora valutano i contenuti sulla base di altri parametri ad esempio, quanto credibile e attendibile sia la fonte che veicola il contenuto, di quale reputazione goda la fonte, la gradevolezza del soggetto che veicola l'informazione. Insomma, tutto ciò che ha a che fare con il confezionamento e con l'aspetto esteriore del contenuto stesso.

Le Fake News non sono un problema che riguarda solo l'informazione, o solo i mezzi di comunicazione, ma più in generale il sistema sociale e l'immaginario collettivo, entrambi divenuti deboli e frammentati.

I social network sono ormai diventati il terreno di coltura e di diffusione ottimale del virus della disinformazione, con conseguenze che vanno ben al di là del recinto del mondo digitale.

Le Fake News sono l'evidenza che la rete è manipolabile, tuttavia non vengono percepite come minaccia, tanto che per l'84,80% le Fake News non devono essere proibite con una legge.

La possibilità che la rete e i social media offrono è quella di comunicare senza limiti, di informarsi senza limiti, di conoscere senza limiti: è questa la presunta libertà percepita dall'utente.



## IMPATTO PERCEPITO DEL SISTEMA COMPLESSIVO DI FAKE NEWS



SONO  
CIRCONDATO  
DA FAKE NEWS



IL SISTEMA  
DEI MEDIA È INVASO  
DA FAKE NEWS



SONO L'EVIDENZA  
CHE LA RETE  
È MANIPOLABILE



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



ESISTONO DA SEMPRE  
MA CON I SOCIAL  
SONO PIÙ EVIDENTI



SONO GENERATE DA  
TUTTI I MEDIA  
E NON SOLO DALLA RETE



NON MI COLPISCONO

## IMPATTO PERCEPITO DELLE FAKE NEWS SUL SISTEMA POLITICO



NON HANNO  
CONDIZIONATO LE  
ELEZIONI POLITICHE  
2018



SONO UNA NUOVA  
FORMA DI PROPAGANDA  
POLITICA



INDEBOLISCONO  
LA DEMOCRAZIA

## CAPITOLO 4

### CLASSIFICAZIONE DELLE NOTIZIE

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

Innanzitutto, per costruire un dataset valido è necessario avere un insieme di articoli direttamente classificati in categorie (es. Reali vs Fake, Reali vs Satiriche vs Clickbait vs Propaganda).

Risulta inoltre necessaria una ripartizione in livelli di analisi (a livello di frase o di documento), partendo da frasi pre-categorizzate manualmente su una varietà di argomenti, classificate in una scala crescente di verità ('bufala', 'falsa', 'a malapena vera', 'mezza verità', 'per la maggior parte vera', 'vera').

Imparare a classificare le fonti come notizie false o reali risulta facile, ma l'obiettivo in questo caso è farlo a partire da tipi specifici di linguaggi e modelli linguistici, focalizzando l'apprendimento su qualcosa di più significativo e generalizzabile.

Si rimuovono inoltre autore, fonte, data, titolo e qualunque altro elemento che possa legare l'apprendimento al dataset scelto, improntando sulle caratteristiche intrinseche di una news reale o fake.

Di seguito si elencano i sei dataset utilizzati per creare il database Fake News Net, sopra citato:

- **BuzzFeedNews:** dataset che comprende un totale di 1672 articoli pubblicati su Facebook da 9 agenzie di stampa diverse e selezionati da 5 giornalisti del portale BuzzFeed.
- **LIAR:** dataset ottenuto usando l'API di PolitiFact che contiene oltre 12836 citazioni di persone ottenute da diversi mezzi, come ad esempio comunicati stampa, interviste televisive etc.
- **BS Detector:** dataset ottenuto tramite l'utilizzo dell'estensione per browser chiamata BS Detector, sviluppato per verificare la veridicità della notizia.
- **CREDBANK:** dataset contenente più di 60 milioni di tweet (inerenti a circa 1000 notizie di eventi differenti) raccolti in una finestra temporale di circa 96 giorni a partire dagli inizi di ottobre 2015 e valutati da circa 30 commentatori di Amazon Mechanical Turk.
- **BuzzFace:** dataset ottenuto estendendo il dataset BuzzFeedNews, includendo i commenti relativi alle notizie raccolte su Facebook. Il set contiene 2263 articoli e 1.6 milioni di commenti.
- **FacebookHoax:** dataset contenente circa 15.500 post di Facebook presi da 32 pagine differenti, tramite l'apposita API. I post sono relativi a notizie scientifiche ed a bufale cospiratrici.

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  $\epsilon$  di notizie predice se esse sono fake news oppure no:

$$F : \epsilon \rightarrow \{0, 1\}$$

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

Partendo da questa definizione binaria del problema (ovvero, una notizia è false oppure no), sono state definite delle classificazioni che prevedono se una notizia è fake oppure no:

- **True Positive (TP):** quando una notizia predetta come Fake News si rivela tale
- **True Negative (TN):** quando una notizia predetta come autentica si rivela tale
- **False Negative (FN):** quando una notizia predetta come autentica si rivela essere una Fake News.
- **False Positive (FP):** quando una notizia predetta come Fake News si rivela essere una notizia autentica.

Definiti i tipi di classificazione per le notizie raccolte, è possibile definire le seguenti metriche che ne permettono lo studio:

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

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

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

$$\text{F}_1 \text{ Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

La **precision** misura il rapporto fra le Fake News individuate e tutte quelle predette come tali. Questo parametro però, per grossi dataset, potrebbe risultare molto elevato alterando così la valutazione di predizione. Il **recall** viene usato per misurare la sensibilità o il rapporto tra le Fake News predette e quelle rivelate come tali.

L'**accuracy** misura la similarità tra le Fake News predette e quelle reali.

L'**F1 Score** è usato per combinare la precision e il recall, permettendo di dare una valutazione complessiva sul rilevamento di Fake News. Più sono alti questi quattro valori, migliore è il rilevamento.

## CAPITOLO 5

### L'INTELLIGENZA ARTIFICIALE APPLICATA ALLE FAKE NEWS

Gli umani non sono molto bravi a identificare le Fake news, pertanto è lecito porsi la seguente domanda: le macchine possono fare o meno un lavoro migliore?

Le macchine sono certamente migliori nel calcolare e tenere traccia dei dati statistici (es. Scoprire i verbi più usati in un testo). Inoltre, possono essere più efficienti nel rilevamento di una base di conoscenza per ritrovare articoli o risposte pertinenti basandosi su più fonti.

Dunque, l'intelligenza artificiale può dare un enorme aiuto nella rilevazione di notizie false diffuse via web, imparando ad analizzare i dati e a trovare modelli per produrre rapidamente risultati che gli umani non potrebbero mai conseguire.

#### Un parallelismo con gli Spam detection

Il problema di rilevare fonti di informazione non genuine basate sull'analisi del contenuto è considerato risolvibile, almeno nell'ambito dello Spam detection, utilizzando tecniche statistiche di machine learning, un ramo dell'intelligenza artificiale, per classificare il testo (tweets o e-mail) come spam.

Il compito di individuare le Fake News è simile, e per la maggior parte analogo, a quello di individuare lo spam, poiché l'obiettivo di entrambi è separare esempi di testo legittimo da quelli di testo illegittimo o con scopi malevoli.

Al posto di usare un filtraggio, come per lo spam, sarebbe utile poter contrassegnare gli articoli come potenziali Fake News, così da poter avvisare i lettori che ciò che si sta leggendo è probabilmente una notizia falsa o ingannevole.

Il rilevamento di notizie false, a differenza del rilevamento di spam, ha molte sfumature che non sono facilmente rilevabili dall'analisi del testo; ad esempio, gli umani hanno bisogno di applicare le loro conoscenze su un particolare argomento per decidere se quella news è vera o falsa. La "legittimità" di un articolo può cambiare semplicemente rimpiazzando un nome di persona con un altro, perciò il meglio che si può ricavare da un approccio content based è decidere se un articolo richiede o meno ulteriore esame.

#### Uso del Machine Learning per il rilevamento di Fake News

Il Machine Learning si occupa di identificare meccanismi che permettono a una macchina intelligente di migliorare le proprie capacità e prestazioni nel tempo. La macchina sarà in grado di imparare a svolgere determinati compiti migliorando, tramite l'esperienza, le proprie capacità, le proprie risposte e funzioni.

Nel campo del Machine Learning, si sono poi sviluppati dei modelli computazionali detti "reti neurali" utili per la generazione di testo, oggi sfruttate anche per generare false notizie.

Nell'ambito delle Fake News, il Machine Learning, dunque, può avere un duplice ruolo. Esso può essere utilizzato sia per generare articoli falsi, secondo il concetto delle "Neural Fake News", ovvero Fake News auto-generate da AI, sia per scovare notizie false generate automaticamente o scritte da un autore umano.

Di seguito, riportiamo una serie di studi e soluzioni che hanno ottenuto maggior riscontro nel campo della rilevazione e identificazione di Fake News.

#### IL MODELLO OPENAI GPT-2

Negli ultimi anni, diversi framework di Natural Language Processing (NLP), un'area dell'informatica e dell'intelligenza artificiale che si occupa delle interazioni fra computer e linguaggi umani, e di come programmare i computer al fine di processare correttamente grandi quantità di dati in linguaggio naturale, sono stati utilizzati per creare e diffondere informazioni false. Il fatto che alcuni di questi modelli di NLP, già predisposti, come BERT, GPT-2, etc. siano di facile utilizzo e a disposizione di tutti, aumenta di molto il rischio che essi vengano utilizzati a scopi propagandistici.

Una Fake News acquista credibilità imitando la forma e lo stile di notizie autentiche, in modo tale da rendere molto difficile la distinzione tra notizia vera e notizia falsa.



Il modello OpenAI GPT-2, ad esempio, 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 articolo. Considerata la natura molto potente di sistemi come il GPT-2 e il suo potenziale uso in attività di disinformazione, OpenAI non ha rilasciato il modello completo del sistema, ma solo una sua versione ridotta. Di seguito, un esempio di testo generato da OpenAI GPT-2:

SYSTEM PROMPT (HUMAN-WRITTEN)	<i>Legolas and Gimli advanced on the orcs, raising their weapons with a harrowing war cry.</i>
MODEL COMPLETION (MACHINE-WRITTEN, FIRST TRY)	<p>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.</p> <p>"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!"</p> <p>"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!"</p> <p>"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."</p>

Nello specifico il modello originale è stato addestrato con 40 GB di dati provenienti da internet e possiede circa 1.5 miliardi di parametri. Invece, il modello ridotto, che è stato rilasciato, ha 117 milioni di parametri. Inoltre, si può pensare di addestrare un modello NLP per cercare di capire se un testo è stato generato automaticamente. Google e Facebook hanno sviluppato i loro modelli di detection, rispettivamente BERT e RoBERTa. Sebbene i due modelli siano simili, e non solo nei nomi, il secondo è un miglioramento del primo che ha mostrato un'accuratezza nell'identificare una Fake News generata automaticamente del 95%.

#### **L'ALGORITMO DELL'UNIVERSITÀ DEL MICHIGAN**

I ricercatori dell'Università del Michigan e dell'Università di Amsterdam hanno lavorato allo sviluppo di uno strumento di intelligenza artificiale basato sull'apprendimento automatico per rilevare le notizie false.

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.

Questa intelligenza artificiale è diversa da un algoritmo di fact checking che controlla le informazioni contenute in un articolo confrontandole con altre fonti al fine di vedere se l'articolo esaminato contiene informazioni incoerenti. La nuova soluzione basata sull'apprendimento automatico potrebbe automatizzare completamente il processo di rilevamento.

Un algoritmo per rilevare le Fake News deve essere in grado di riconoscere le notizie non false, verificando la verità e tenendo conto di fattori come lo sviluppo delle notizie e le interpretazioni linguistiche e culturali. I ricercatori hanno creato il proprio set di dati chiedendo agli operai di Amazon Mechanical Turk di reinterpretare 500 notizie reali come fossero Fake News. Ai partecipanti allo studio è stato chiesto di imitare lo stile giornalistico dell'articolo originale, ma gonfiando i fatti e le informazioni per garantire che il risultato fosse chiaramente falso.

Il gruppo di ricercatori ha poi fornito all'algoritmo sia le Fake News che le notizie vere in modo che imparasse a distinguere le due cose.

Dopo questa prima fase di apprendimento, il sistema è stato messo alla prova con una serie di notizie vere e false prese direttamente dal Web. L'algoritmo è riuscito a individuare le bufale nel 76% dei casi, mentre l'occhio umano normalmente si ferma al 70%. Inoltre, ha dimostrato di essere applicabile a notizie di qualsiasi argomento e di poter operare più velocemente dell'uomo: riesce infatti a bloccare le bufale non appena compaiono sulla rete, prima ancora che sia possibile verificarle attraverso controlli incrociati con altri testi.

Quel che è certo, è che un accurato fact checking, per quanto doveroso, ha bisogno di tempo, capita molto spesso che durante il tempo necessario per capire se una notizia è vera o è falsa, la stessa sia già stata diffusa su larga scala. Inoltre, ci si potrebbe anche chiedere "chi controlla il controllore?", infatti molti siti di Fact Checking, potrebbero avere essi stessi un bias di natura politica e discriminare le notizie in base a questo.

#### LA SOLUZIONE PROPOSTA DAL LABORATORIO CSAIL

Un'altra possibile soluzione al problema, proposta dal laboratorio CSAIL (Computer Science & Artificial Intelligence Lab) del MIT, è quella di 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.

Questi ricercatori hanno utilizzato vari metodi per analizzare i siti dei media, gli account twitter associati, la reputazione della fonte, il traffico web e altri fattori, al fine di ipotizzare classifiche di alta, media e bassa veridicità.

Dai dati elaborati dal loro algoritmo emerge che la struttura di testo degli URL di una fonte è significativa per la veridicità: gli URL che hanno molti caratteri speciali e sottodirectory complicate, per esempio, sono associati a fonti meno affidabili. Se un sito web ha già pubblicato notizie false, ci sono buone probabilità che lo faccia di nuovo. Il sistema ha bisogno solamente di circa 150 articoli per rilevare in modo affidabile se una fonte di notizie può essere attendibile. Un approccio di questo tipo sembra essere perfetto per aiutare a individuare nuovi punti di spaccio di Fake News prima che le storie si diffondano troppo.

#### FAKE NEWS DETECTION

L'obiettivo principale del Fake News Detecting comprende l'identificazione del linguaggio (come insieme di parole o frasi) utilizzato per ingannare i lettori; questa idea di classificazione per apprendimento è un task impegnativo.

Tipo di fake news	Esempio
100% Falsa	#RIP Paul McCartney
Tendenziosa e di parte	<i>News da fonte A:</i> Il cambiamento climatico produrrà più uragani come l'Uragano Katrina <i>News da fonte B:</i> il cambiamento climatico può portare a grandi uragani -> non si sono verificati grandi uragani -> il cambiamento climatico non esiste
Uso improprio dei dati	"Bevi birra, fa bene al cervello", riporta Inc. Ma dovrete aspettare un momento prima di farvi una pinta (o due). Lo studio è stato fatto sui topi, non sugli umani. E la quantità di birra era l'equivalente di 28 fusti per una persona.
Imprecisa e sciatta	"1 CEO su 5 è Psicopatico, secondo uno studio". Ma l'headline è sbagliata. La ricerca è basata su un sondaggio sugli operai di una industria a catena di montaggio, non sui CEOs.

Gli esempi qui sopra catturano la natura complessa del task di individuazione e categorizzazione di una Fake News, infatti, per classificare correttamente i suddetti tipi di Fake News il modello linguistico deve comprendere le sottigliezze coinvolte nella trasmissione di messaggi attraverso un testo.

La sola identificazione manuale è già di per sé un compito difficile, complicato dal fatto che oggi le news non provengono più soltanto dai media tradizionali, ma anche tramite diversi canali di social media, per cui una soluzione automatica richiede la comprensione del linguaggio naturale, di sua natura complesso e intricato, il che rende il tutto una bella sfida.

Di grande aiuto è il già citato Natural-Language processing (NLP) che permette di scremare e convertire dati testuali in un formato leggibile dalle macchine.

Diventa quindi necessario utilizzare le parole come vettori (word vector representations or word embedding), di seguito sono elencati alcuni modelli:

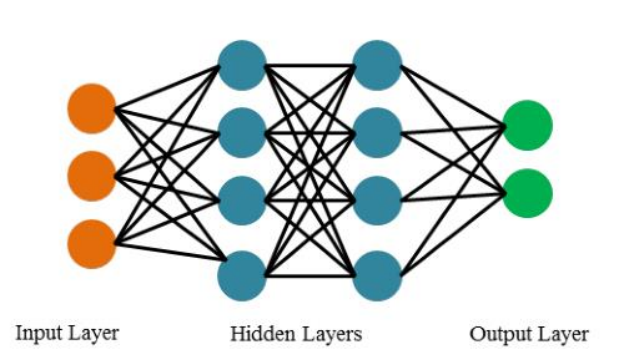
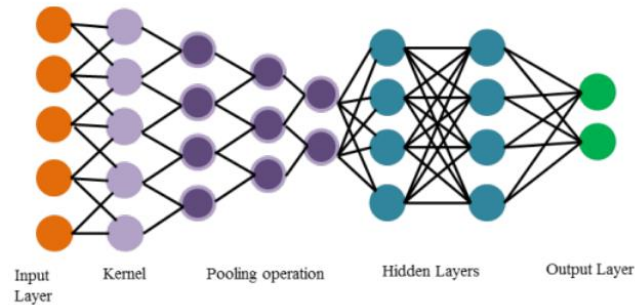
**Bag of words (BoW):** le frasi sono rappresentate come multiset di parole, ignorando l'ordine e il contesto della parola.

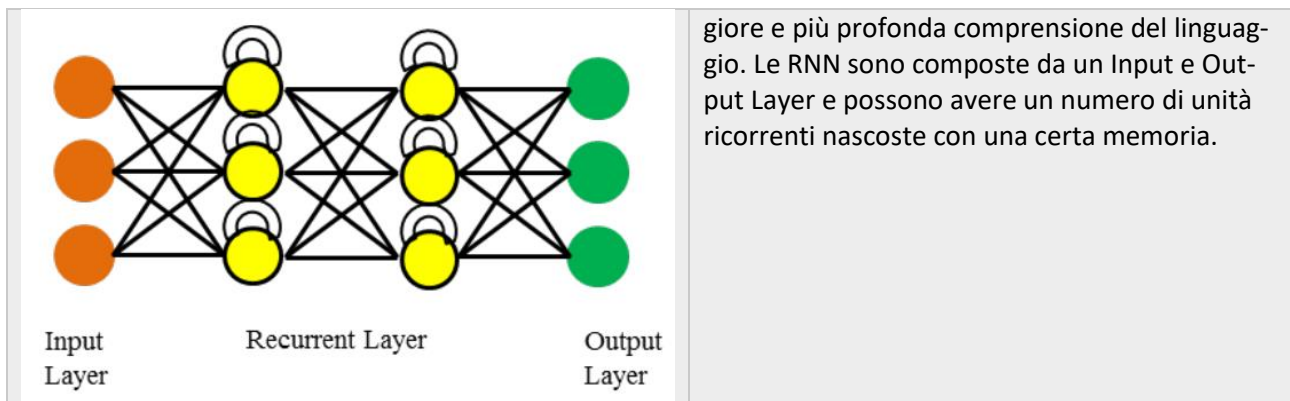
**Tf-Idf (Term Frequency-Inverse Document Frequency):** dà un'indicazione sull'importanza di una data parola per una frase o documento rispetto ad una intera collezione.

**GloVe:** si costruisce una matrice di co-occorrenza con tutte le parole nella collezione, riducendone successivamente le dimensioni con metodi di fattorizzazione.

**Word2Vec:** si tratta di un modello predittivo disegnato per la predizione di parole data una finestra di contesto.

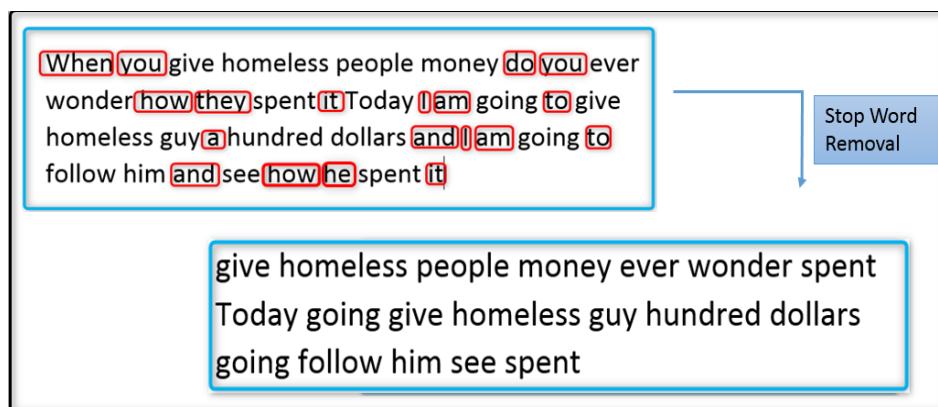
Cruciale è anche la scelta del tipo di rete neurale da utilizzare, le più comuni sono:

Model Name	Model Description
<b>Dense Neural Network (DNN)</b>  <p>Input Layer      Hidden Layers      Output Layer</p>	<p>La rete densa e completamente connessa consente di passare l'input come sequenza di parole. L'architettura a strati permette di sperimentare la giusta profondità che serve per il task. La rete è costituita da un Input Layer, un Output Layer intervallati da una serie di Hidden Layers.</p>
<b>Convolutional Neural Network (CNN)</b>  <p>Input Layer      Kernel      Pooling operation      Hidden Layers      Output Layer</p>	<p>Le CNN sono molto simili alle normali Reti Neurali: sono costituite da neuroni che hanno pesi e influenze apprendibili. La CNN è composta da un Input e Output Layer e da più Hidden Layers, i quali solitamente consistono di Convolutional Layers, Pooling Layers, Fully Connected Layers e Normalization Layers.</p>
<b>Recurrent Neural Network (RNN)</b>	<p>Le RNN sono popolari avendo a che fare con dati sequenziali in cui ogni unità può memorizzare lo stato dell'unità precedente. Ciò è particolarmente utile nell'elaborazione del linguaggio naturale perché aiuta ad ottenere una mag-</p>

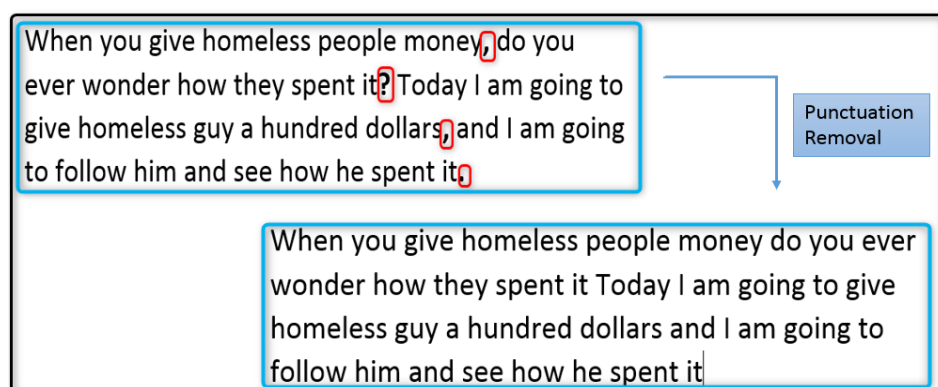


I dati testuali richiedono inoltre uno speciale pre-processing del testo per poter implementare algoritmi di machine learning o deep learning su di essi; esistono varie tecniche per convertire il testo in una forma pronta per il modeling, che consistono in passi sequenziali:

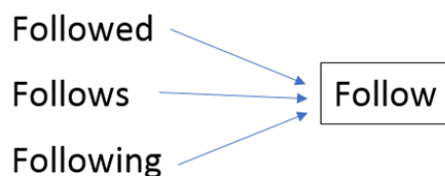
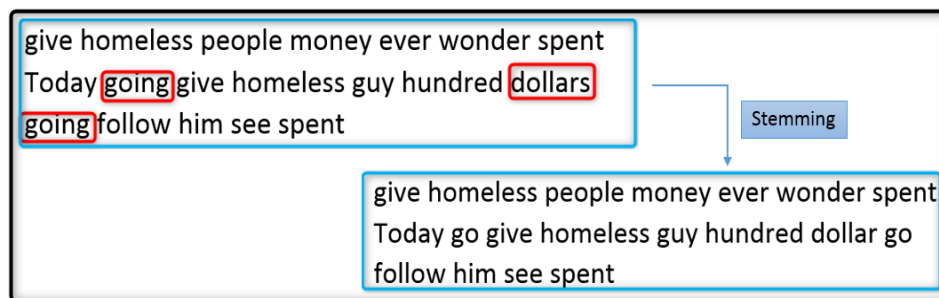
- **Rimozione delle Stop Word:** si inizia rimuovendo le stop words (le parole più comuni in un linguaggio che non forniscono contesto, come congiunzioni, preposizioni e articoli) dal testo



- **Rimozione della Punteggiatura:** la punteggiatura non è di gran valore nel comprendere il significato di una frase



- **Stemming:** si tratta di una tecnica che rimuove prefissi e suffissi da una parola, lasciandone la radice; usando lo stemming si possono ridurre le forme flesse e talvolta le forme derivate di una parola ad una forma base comune.



- **Word Vector Representation:** per eseguire l'analisi del testo bisogna convertire il testo grezzo in features numeriche, per effettuare questa trasformazione è possibile utilizzare due tecniche: Bag of Words e TF-IDF.

La tecnica di BoW elabora ogni articolo di notizie come un documento, e ne calcola la frequenza di ogni parola, utilizzata per creare una rappresentazione numerica dei dati (vector features of fixed length). Questa metodologia presenta degli svantaggi in termini di perdita di informazioni, non tenendo traccia della posizione delle parole o sul contesto.

La Term Frequency identifica l'importanza locale di una parola in base alla sua presenza in un documento, mentre l'Inverse Document Frequency identifica le signature words, che non compaiono spesso nei documenti. Una parola con un alto TF-IDF è una signature word per quel documento, avendo alta frequenza nel documento ma non comune nella collezione.

## CAPITOLO 6

### PRINCIPALI TOOL PER L'IDENTIFICAZIONE DI FAKE NEWS

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

#### Grover

I ricercatori dell'Università di Washington e dell'Allen Institute hanno sviluppato un modello di AI che riesce a identificare le Fake News con una precisione del 92%, secondo i test effettuati dal team, un incremento sostanziale rispetto ad altri tool di fake news detection, il migliore dei quali riesce a raggiungere un'accuratezza del 73%.

L'algoritmo utilizzato, soprannominato Grover (Generating aRticles by Only Viewing mEtadata Records), riesce ad analizzare diversi aspetti di un articolo, inclusi il corpo, l'intestazione, il nome dell'autore, il nome della pubblicazione ed altri dettagli che possono indicarne l'infondatezza.

Grover è stato addestrato con una libreria di 120GB di articoli attendibili tratti dalle prime 5000 pubblicazioni individuate da Google News da fine 2016 a metà 2019; ciò che lo rende così efficace nello scovare contenuti non affidabili è il suo essere anche in grado di crearne: data un'intestazione da cui partire l'algoritmo può generare un intero articolo notiziario scritto con lo stile di una redazione o di uno specifico autore.

#### Botometer

Botometer è un algoritmo di apprendimento automatico addestrato per classificare un account sul social network Twitter come bot o umano basato su decine di migliaia di esempi etichettati.

Nel controllo, il browser recupera il profilo pubblico da esaminare e centinaia dei suoi tweet e menzioni pubblici utilizzando l'API di Twitter, questi dati vengono passati all'API Botometer, che estrae circa 1.200 feature per caratterizzare il profilo dell'account, come gli amici, la sua struttura sul social network, i pattern di attività nel tempo, la lingua e l'insieme di opinioni. Infine, le feature sono utilizzate da vari modelli di apprendimento automatico per calcolare 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.

Data la forte dipendenza dalla lingua di alcune feature il bot è destinato ad essere usato su account in lingua inglese, ma possono in ogni modo essere utilizzate le restanti per produrre un'analisi indipendente dalla lingua.

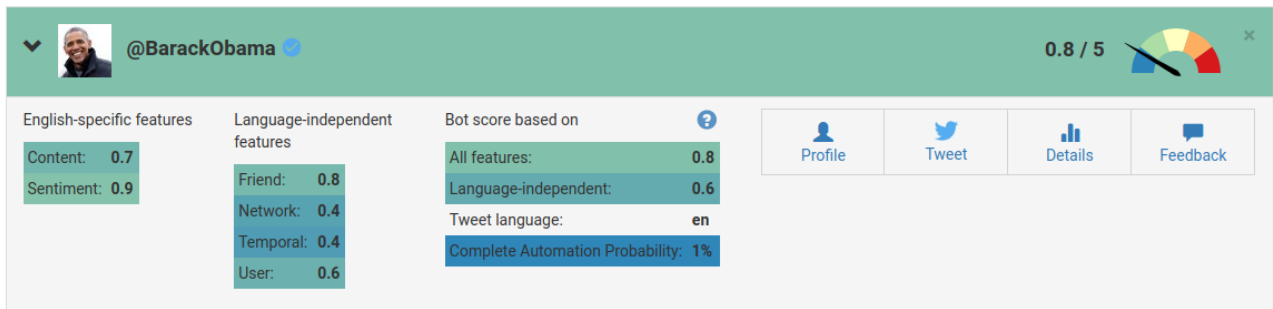
È possibile utilizzare il tool in due modalità: tramite il sito web *botometer.iuni.iu.edu* oppure tramite le API Python messe a disposizione dagli sviluppatori. La versione web presenta una home in cui si deve dare il permesso al tool di accedere al proprio profilo Twitter: questo passaggio è necessario affinché il tool possa usare le API messe a disposizione da Twitter per l'analisi dei profili. Una volta dati i permessi necessari si può procedere con l'analizzare uno specifico profilo oppure l'intera lista di follower.

Il bot esegue un controllo su tutti i tweet inviati, i retweet effettuati e i "mi piace" lasciati a tweet di altri utenti. L'analisi di basa su tre categorie di caratteristiche:

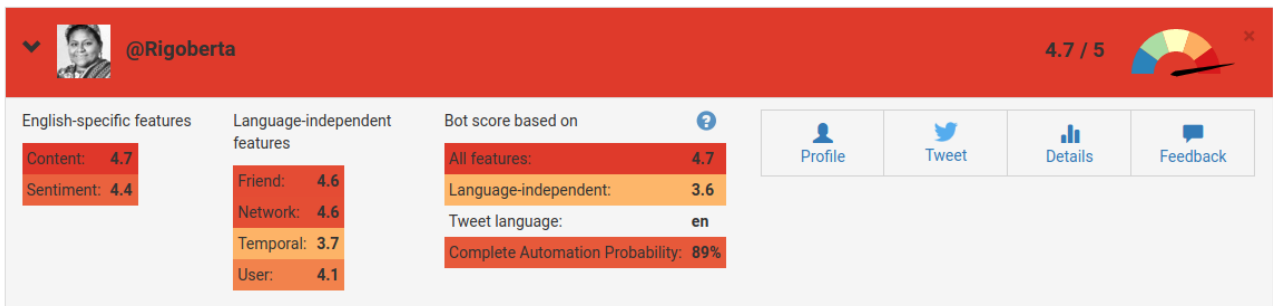
- **Basate sull'inglese:** che analizzano il contenuto e il sentimento espresso dai tweet.
- **Indipendenti dal linguaggio:** che analizza l'intero profilo utente (amici, timeline, etc.).
- **CAP – Complete Automation Probability:** indica quanto è probabile che il profilo analizzato sia un bot.

Il tool darà per ogni parametro un valore da 0 a 5. I valori intermedi tra il minimo e massimo possono indicare incertezza sull'effettiva valutazione data ad un profilo. Inoltre, il valore CAP non assicura la totale certezza che un profilo sia completamente automatizzato e dunque un bot.

Analisi di un profilo Twitter autentico:



Analisi di un profilo Twitter gestito da un bot:



Le API Python svolgono le medesime analisi, con la differenza che tutte le operazioni devono essere lanciate da terminale utente.

Ecco un esempio di codice utilizzato per analizzare un account Twitter:

```
import botometer

mashape_key = "xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx"
twitter_app_auth = {
    'consumer_key': 'xxxxxxxx',
    'consumer_secret': 'xxxxxxxx',
    'access_token': 'xxxxxxxx',
    'access_token_secret': 'xxxxxxxx',
}

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)

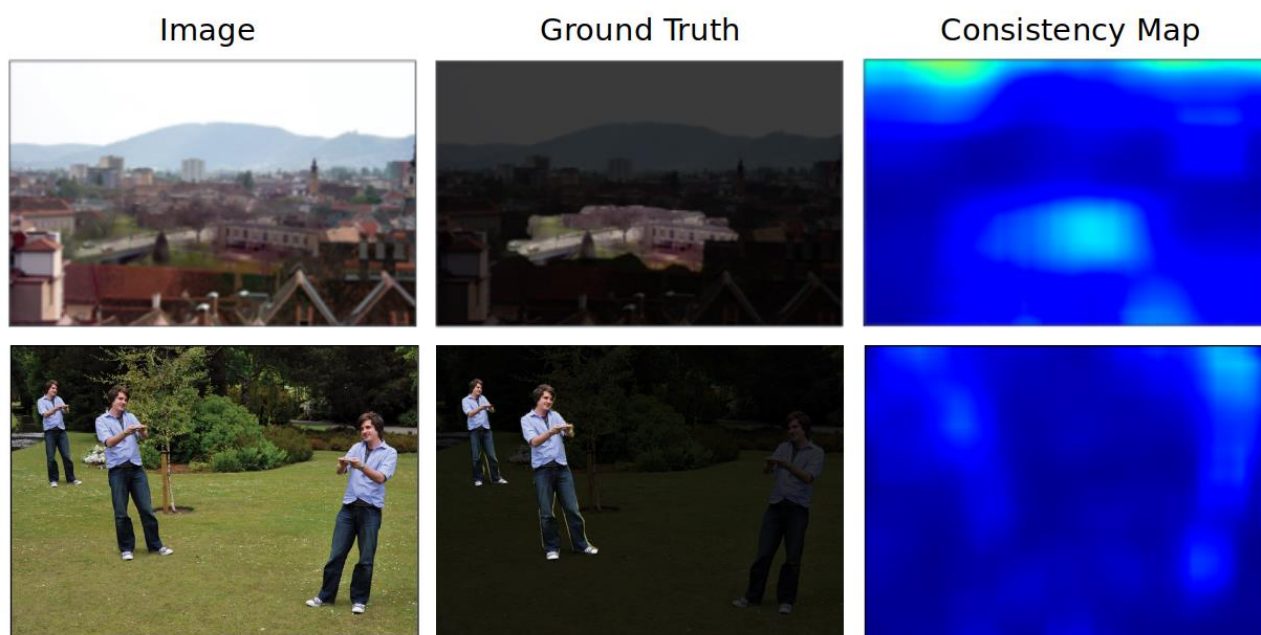
# Check a sequence of accounts
accounts = ['@clayadavis', '@onurvarol', '@jabawack']
for screen_name, result in bom.check_accounts_in(accounts):
    # Do stuff with `screen_name` and `result`
```



### Image Self-Consistency tool

I progressi negli strumenti di fotoritocco e manipolazione hanno reso notevolmente più semplice la creazione di immagini false. Imparare a rilevare tali manipolazioni, tuttavia, rimane un problema difficile a causa della mancanza di quantità sufficienti di dati di addestramento manipolati.

Questo tool utilizza un algoritmo di apprendimento per il rilevamento di manipolazioni di immagini visive che viene addestrato utilizzando solo un ampio set di dati di fotografie reali. L'algoritmo utilizza i metadati EXIF fotografici registrati automaticamente per addestrare un modello per determinare se un'immagine è autoconsistente, ovvero se il suo contenuto avrebbe potuto essere prodotto da una singola imaging pipeline. 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.



### Fakebox

Fakebox analizza articoli di notizie per valutare se è probabile che siano o meno notizie reali. Osservando una serie di aspetti disponibili di un articolo (titolo, contenuto e URL) utilizzando modelli di apprendimento automatico integrati e un database a cura manuale, Fakebox è in grado di identificare con successo le notizie false e analizza gli articoli per valutare se è probabile che siano notizie vere o no.

Fakebox controlla i seguenti aspetti di un articolo. Gli utenti devono fornire almeno una proprietà da controllare, ma più informazioni vengono fornite su Fakebox, più utili saranno i risultati.

**Titolo:** i titoli possono essere clickbait o distorti.

**Contenuto:** Il contenuto testuale di un articolo può essere analizzato per determinare se è scritto come una vera notizia o no.

**Nome dominio:** Alcuni domini sono noti per ospitare determinati tipi di contenuti, Fakebox conosce i siti più popolari.



## CAPITOLO 7

### FAKE NEWS E COVID-19

Durante l'emergenza da Coronavirus, al rischio sanitario si unisce pericolosamente il rischio sociale e della fiducia alimentato da vere e proprie pandemie informative che "infettano" uno dei più importanti fattori della gestione della crisi: la corretta comunicazione dell'emergenza.

Il risultato è una grande confusione, in un momento in cui – al contrario – il nostro bisogno di notizie chiare è alle stelle. I rischi sono molto concreti: possiamo mettere in atto comportamenti sbagliati, perdere fiducia nelle indicazioni ufficiali, non capire le restrizioni imposte.

Il 4 aprile Palazzo Chigi ha deciso di istituire una 'task force' anti-fake news, più propriamente un'unità di monitoraggio contro la diffusione di notizie false relative al Covid-19 sul web e sui social network. L'unità, messa in piedi dal sottosegretario con delega all'Editoria Andrea Martella, è composta da rappresentanti del Ministero della Salute, della Protezione civile che lavorano insieme ad esperti provenienti dal mondo accademico e del giornalismo.

L'obiettivo del nuovo organismo è analizzare lo scenario informativo sul virus e suggerire strumenti o fornire indicazioni utili alle istituzioni per comunicare in modo appropriato e veicolare informazioni di qualità. Ad esempio, rendendo più efficace la comunicazione su Covid-19, prendendo in considerazione i bisogni dei cittadini e le tematiche attorno alle quali si genera più confusione.

L'OMS (Organizzazione Mondiale della Sanità), invece, ha direttamente chiesto la collaborazione di Google, Apple, Twitter, Facebook, Instagram, TikTok, etc. per filtrare le informazioni false e promuovere informazioni accurate da fonti credibili.

Le piattaforme digitali hanno risposto all'appello dell'OMS applicando di fatto varie restrizioni indirizzate contro le applicazioni ritenute non affidabili perché non certificate da enti istituzionali o di ricerca riconosciuti. In sostanza, per combattere il diluvio di disinformazione, si sta costruendo una banda di rivelatori di verità che disperdono i fatti e sfatano i miti.

Ma vediamo nel dettaglio le strategie che le principali piattaforme social hanno adottato per contrastare la crescente disinformazione:

- **Twitter** sceglie di concentrarsi su fonti ufficiali quali l'OMS e il CDC nei risultati di ricerca tramite il termine "coronavirus". In un post di fine gennaio Twitter ha riferito di aver assunto provvedimenti per garantire che la piattaforma fosse protetta da "comportamenti dannosi" e che avrebbe rimosso gli account coinvolti in tali comportamenti.
- Per quanto riguarda **Facebook**, quando le persone cercano informazioni relative al virus si attiverà un pop-up educativo con informazioni attendibili multilingue.
- Stessa logica per **Instagram** che sta bloccando o limitando gli hashtag utilizzati per diffondere la disinformazione attraverso popup mirato che in caso di ricerche basate su termini come "coronavirus" riportano alla pagina informativa dell'Organizzazione mondiale della sanità.

WhatsApp, Signal, iMessage e Telegram, essendo crittografati end-to-end, invece, mal si prestano a strategie mirate contro i contenuti Fake virali.

Nonostante WhatsApp permetta solo di condividere messaggi privatamente ed è ben distante dal grande pubblico di Facebook, Instagram o Twitter, risulta essere il luogo virtuale più pericoloso per la diffusione di notizie false. Il motivo è che la portata di una delle applicazioni più scaricate al mondo è enorme, e il fatto che tramite WhatsApp si possano inviare o inoltrare messaggi in via totalmente privata, rende ancora più difficile la lotta alla loro diffusione.

Le preziose funzionalità di sicurezza di WhatsApp impediscono all'azienda stessa di moderare i contenuti condivisi nella piattaforma e certo, l'aggressività di moderazione tipica dei social network di quest'ultimo periodo, su WhatsApp non è assolutamente possibile.

Tuttavia, spinti da una pressione globale non indifferente, gli sviluppatori dell'app per i messaggi gratis hanno nel tempo migliorato le "difese digitali" della piattaforma, ad esempio riducendo il numero di volte che i messaggi possono essere inoltrati agli altri e tentando di rallentare la diffusione di tali contenuti. Oltre a questo accorgimento, sono state implementate nuove funzionalità. Ad esempio, l'applicazione permette ai suoi utenti di 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.

Inoltre, basterà 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.

Non solo: gli utenti che lo richiederanno potranno anche ricevere, via WhatsApp, un resoconto giornaliero di tutte le notizie verificate da Facta.

Insomma, sono proprio le caratteristiche che rendono WhatsApp così irresistibile per le persone che in realtà potrebbero contribuire all'aumentare delle probabilità che la piattaforma per i messaggi gratis diventi un terreno fertile per la disinformazione e le Fake News. E come in ogni cosa, anche in questo caso c'è una sola arma in grado di contrastare realmente questo pericoloso fenomeno: il buonsenso.

Una soluzione più efficace potrebbe venire direttamente dai Governi interessati e dai media, che dovrebbero spiegare sempre meglio e con maggiore costanza l'importanza di informarsi esclusivamente tramite gli organi istituzionali ufficiali.

Infatti, a nostro modo di vedere, la soluzione non può essere quella di limitare la circolazione delle informazioni, ma si deve intervenire sulla qualità dei contenuti. Inoltre, delegare ai gestori delle piattaforme digitali il controllo di tale qualità significherebbe dare loro un potere enorme: chi stabilisce cosa è Fake News o cosa no? Piuttosto, sono le istituzioni a dover intervenire, portando sui social network informazioni corrette.

Le istituzioni, in questo caso quelle sanitarie e quelle politiche in particolare, devono imparare a stare di più e meglio sui social media dato che la maggior parte delle persone oggi si informa lì. E invece, a volte ci siamo trovati di fronte all'assenza di propri profili social, anche se Twitter e soprattutto Facebook sono piattaforme che esistono da molti anni. Oltretutto, è chiaro che non ci si può limitare a rispondere a una Fake News con una good news: la prima ha una forza molto maggiore. La questione dunque va affrontata a monte. Un'educazione digitale dovrebbe essere parte strutturale dei programmi di studio, almeno dalle scuole medie inferiori.

## 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.