[CT0540] Social Network Analysis

Year: 2023/2024

Confronto tra Real e Fake news

Lab Project

Brognera Enrico 890406

Abstract

In un'epoca in cui i social media dominano il flusso di informazioni, distinguere tra real news e fake news è diventato cruciale. Questo studio esamina le dinamiche di interazione degli utenti con entrambe le tipologie di notizie, focalizzandosi su due argomenti distinti: politici e di gossip. Attraverso l'analisi dei dati raccolti dal dataset FakeNewsNet del 2018 si è cercato di comprendere il livello di engagement delle notizie e la polarizzazione degli utenti. I risultati suggeriscono una maggiore tendenza all'engagement con le fake news e una marcata polarizzazione basata sulla veridicità delle notizie. Questa ricerca offre spunti significativi per comprendere al meglio come gli utenti interagiscano con notizie differenti e offre uno spunto per future indagini sull'impatto delle fake news nella società digitale odierna.

Index

1 Introduzione **1**

2 Raccolta dei dati **2**
2.1 Tecnologie utilizzate 2

3 Analisi dei dati **3**
3.1 Word frequency 3
3.2 Numero interazioni 5
3.3 Grafo engagement 7
3.4 Affidabilità utenti 8
3.5 Rete dei follower 9

4 Risultati e conclusioni **11**
4.1 Futuri sviluppi 11

1 Introduzione

Le fake news, o notizie false, sono informazioni non verificate, basate su dati inesatti o distorti, con l'obiettivo di ingannare e talvolta manipolare il lettore. Questo fenomeno ha preso piede soprattutto con l'avvento dei social media, dove le notizie possono essere diffuse rapidamente e su larga scala senza un controllo rigoroso della loro autenticità. La forza delle fake news risiede nella loro abilità di sembrare credibili e intriganti, riuscendo così a catturare l'attenzione dei destinatari.[1]

Le conseguenze di questo fenomeno sono molteplici. Le fake news possono influenzare i risultati delle elezioni politiche, generare panico o disinformazione in situazioni di crisi, e danneggiare la reputazione di individui o organizzazioni. Inoltre, possono alimentare la polarizzazione sociale, dato che le persone tendono a circondarsi di notizie che rafforzano le loro convinzioni, ignorando quelle che potrebbero metterle in discussione.[2]

In questo documento cercheremo di esaminare le differenze tra le interazioni degli utenti con le fake news e le real news, analizzando in particolare due categorie di notizie: quelle politiche e quelle di gossip. L'obiettivo di questo documento sarà quindi quello di studiare la differenza di *engagement* tra notizie vere e false e di rispondere alla domanda se esiste o meno una polarizzazione tra gli utenti.

2 Raccolta dei dati

E' stata utilizzata una repository di dati chiamato **FakeNewsNet**, la quale include un insieme di dati con contenuti di notizie, contesto sociale e informazioni spaziotemporali. Queste notizie sono state raccolte da due siti web di verifica dei fatti (*fact-cheching*), GossibCop e PolitiFact, che contengono notizie etichettate da giornalisti professionisti e sono classificate come notizie vere o false. Queste notizie possono riguardare due argomenti distinti: politica o gossip. Le notizie provengono da vari periodi, come indicato dalle date di pubblicazione, e le notizie selezionate per questo studio sono state pubblicate tra il 10 marzo 2018 e il 13 giugno 2018. Poiché i siti utilizzati per la verifica dei fatti analizzano le notizie americane, tutte le notizie provengono dagli Stati Uniti.[4]
Per comprendere meglio i risultati ottenuti, è necessario avere un minimo di contesto sociale in America durante il periodo delle notizie. La primavera del 2018 precede le elezioni parlamentari negli Stati Uniti e, insieme alle elezioni governatoriali, costituiscono le cosiddette elezioni di metà mandato (*Midterm Elections*) durante la presidenza di Donald Trump. Durante queste elezioni, sono stati sottoposti al voto tutti i seggi della Camera dei rappresentanti e i 33 seggi del Senato della classe 1.[6]
Le informazioni sulle interazioni degli utenti sono state raccolte utilizzando le API di Twitter, ogni utente e il numero di volte che ha interagito con una notizia specifica attraverso un tweet sono stati salvati. Infine, è stata condotta un'analisi tra gli utenti per verificare le connessioni tra di loro, quindi sono stati registrati tutti i follower di ogni utente che ha interagito con almeno una delle notizie analizzate.[3]

2.1 Tecnologie utilizzate

Per lo sviluppo e la stesura di questo documento sono state utilizzate le seguenti tecnologie:

- ◇ R: utilizzato per l'analisi del dataset mediante l'IDE. (integrated development environment) *R studio (versione 4.4.0)*
- ◇ Python: utilizzato per la riorganizzazione del dataset mediante l'IDE *Pycharm*.
- ◇ Github: utilizzato per salvare l'avanzamento del codice.
- ◇ Latex: utilizzato per la stesura di questo documento mediante la piattaforma *Overleaf*.

Il formato con cui sono salvati i dati sono *.txt* oppure *.csv* per le relazioni, mentre le informazioni sulle notizie (titoli, testi, autori, ..) in formato *.json*.

3 Analisi dei dati

Le varie analisi sono state effettuate per mezzo di codice R.
Per comprendere al meglio le seguenti analisi abbiamo inanzitutto bisogno di conoscere la mole di dati analizzati.

	political news	**gossip news**
numero fake news	120	91
numero real news	120	91
numero utenti totali	23865	15257
numero tweet totali	37259	25240

In questo ducumento, quindi, analizzeremo distintamente le due categorie di notizie poichè gli utenti in comune tra le due sono minimi ed inoltre ci permetterà di confrontarle visualizzando le possibili differenze o analogie.

3.1 Word frequency

Prima di effettuare le analisi sull'engagement è importante capire il contesto delle notizie prese in analisi. Per effettuare ciò abbiamo deciso di studiare la *word frequency*, quindi l'analisi delle parole più utilizzate.Per fare questo abbiamo pulito il dataset da quelle che vengono chiamate stopwords, ovvero tutte le parole comuni che non hanno a che vedere con un particolare argomento specifico quali ad esempio articoli e preposizioni. Dopo aver rimosso queste parole abbiamo quindi proceduto con l'analisi dei 30 termini più utilizzati (l'istogramma) ed di tutti i vocaboli con un numero maggiore di 30 ripetizioni (grafico a nuvola). Abbiamo effetuato queste analisi studiando distintamente le notizie vere e false, divise a loro volte per categoria.

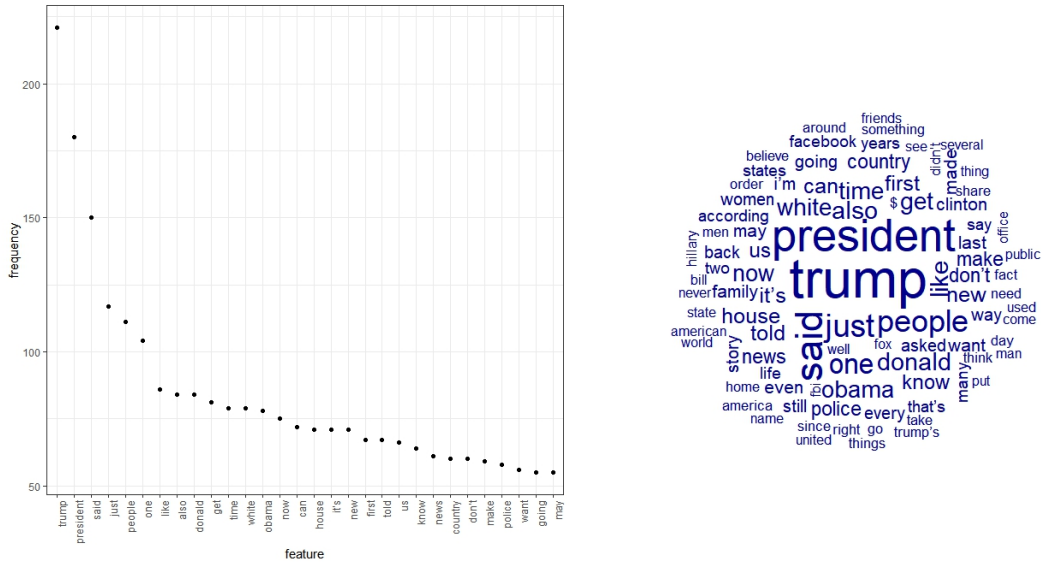

Figure 1: Word frequency fake political-news

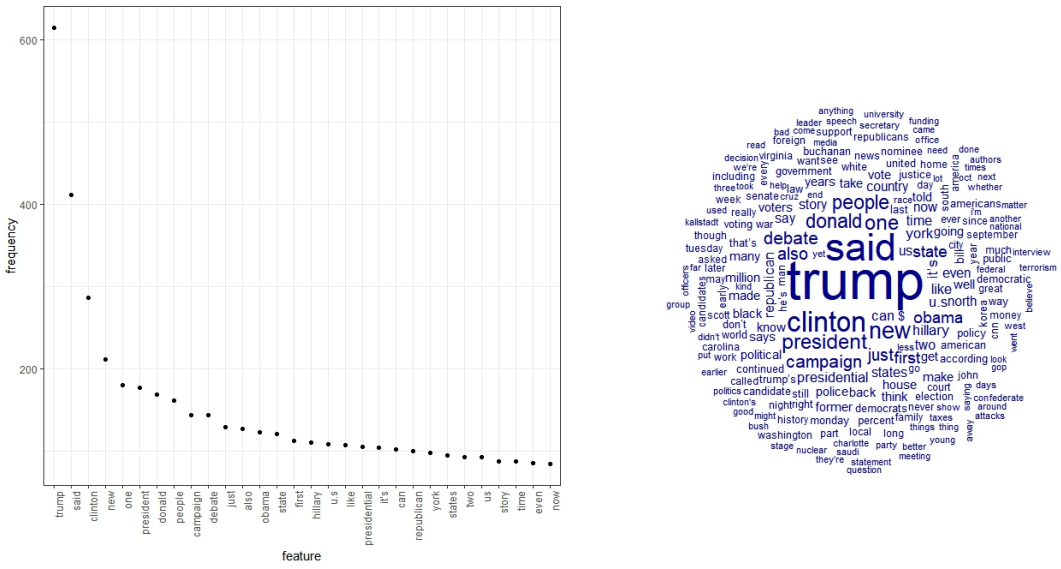

Figure 2: Word frequency real political-news

Possiamo notare nelle figure 1 e 2 come le parole più utilizzate per la categoria di notizie politiche siano abbastanza simili con una forte presenza nel nome del presidente Donald Trump.

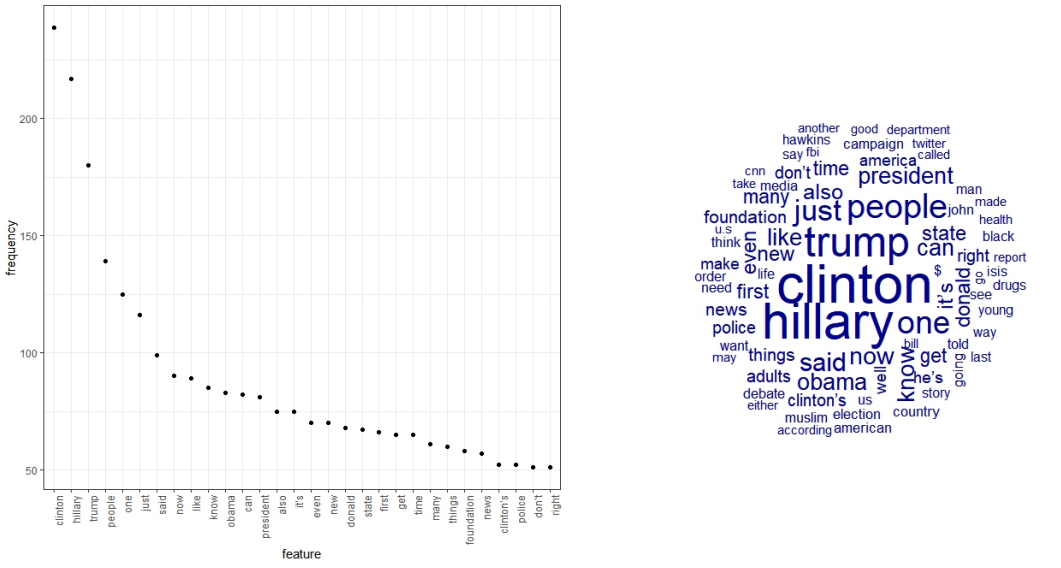

Figure 3: Word frequency fake gossip-news

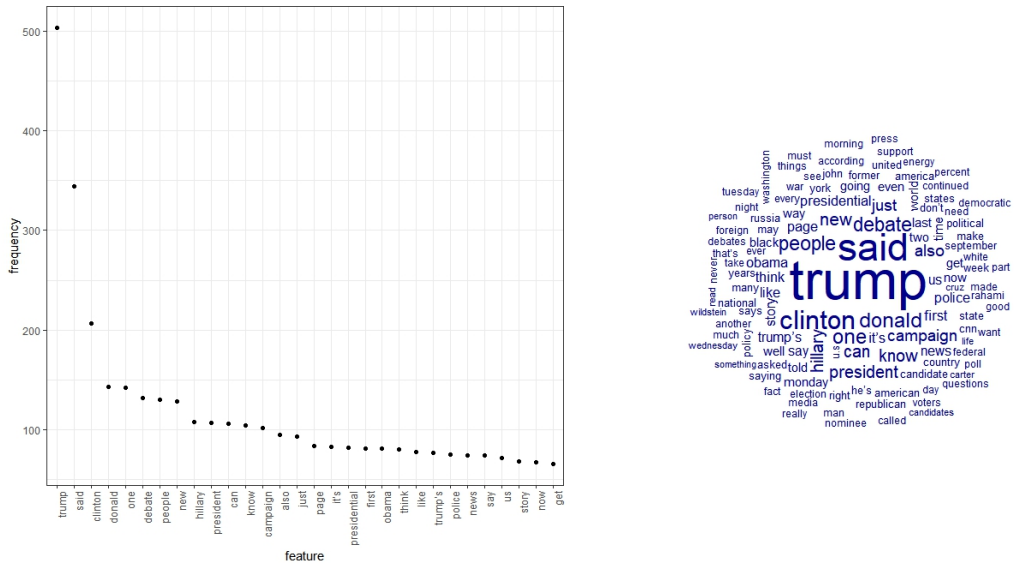

Figure 4: Word frequency real gossip-news

Possiamo notare invece nelle figure 3 e 4 una forte differenza tra le parole utilizzate tra le fake e real news nella categoria gossip. A differenza della categoria precedente nelle feke news c'è una forte presenza del nome Hillary Clinton, principale oppositrice al governo.
In entrambe le categorie possiamo notare come la frequenza delle parole più utilizzare nelle fake news siano minori rispetto alle notizie vere. Questo è determinato dalla lunghezza dei testi delle notizie prese in analisi, poichè nel nostro caso di studio le fake news tendono ad essere più brevi rispetto alle notizie vere.

3.2 Numero interazioni

In questa sezione andremo ad analizzare l'engagement, quindi inizialmente studieremo il numero totale di interazioni delle notizie raggrupandole tra notizie vere e false.

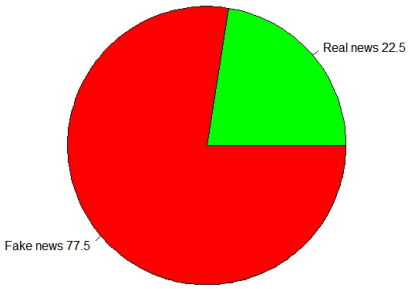

Figure 5: Rapporto tweet real e fake political-news

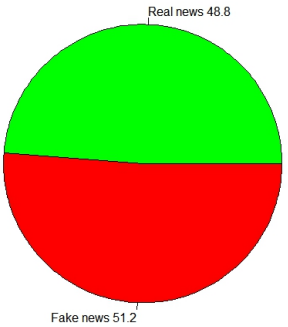

Figure 6: Rapporto tweet real e fake gossip-news

Nelle figure 5 e 6 abbiamo quindi confrontato il numero di tweet inerenti alle fake e real news. É interessante notare come nel nostro campione di notizie politiche la gran parte dei tweet associati appartenagno a quelle false a differenza delle notizie di gossip nelle quali la distribuzione è quasi pari.
Un ulteriore dato importante può essere il numero di utenti univoci che hanno commentato tali notizie.

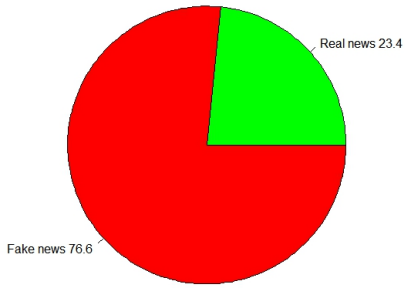

Figure 7: Rapporto utenti real e fake political-news

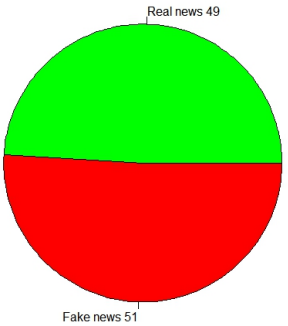

Figure 8: Rapporto utenti real e fake gossip-news

Nelle figure 7 e 8 possiamo notare come le percentuali si discostano leggeremente risepetto a quelle delle figure 5 e 6 notando un leggero decremento delle fake news. Questo ci può portare a dedurre che di media le fake news hanno un maggiore numero di interazioni per singolo utente rispetto alle notizie vere quindi le fake news hanno un maggior grado di

engagement.

3.3 Grafo engagement

La sucessiva analisi è il grafo dell'engagement degli utenti comuni sulle varie notizie. Ciascun nodo del grafo rappresenta una notizia, verde le real-news e rosso le fake-news. Ciascun arco che collega i nodi rappresenta invece un utente che ha interagito con entrambe le notizie. Quindi più archi ha un nodo più tale notizia ha avuto interazioni e più i nodi hanno archi in comune, e quindi sono vicini tra loro, più queste notizie hanno utenti in comune che hanno interagito.
Nota importante: i nodi isolati sono stati eliminati poichè poco interessanti per la nostra analisi siccome non avevano interazioni collegate ad essi.

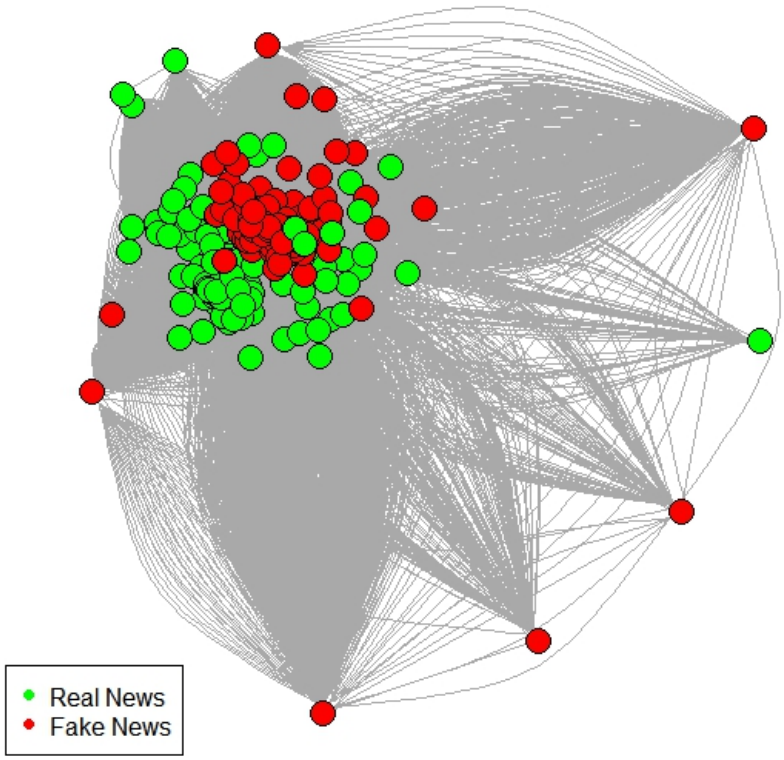

Figure 9: Grafo engagement political-news

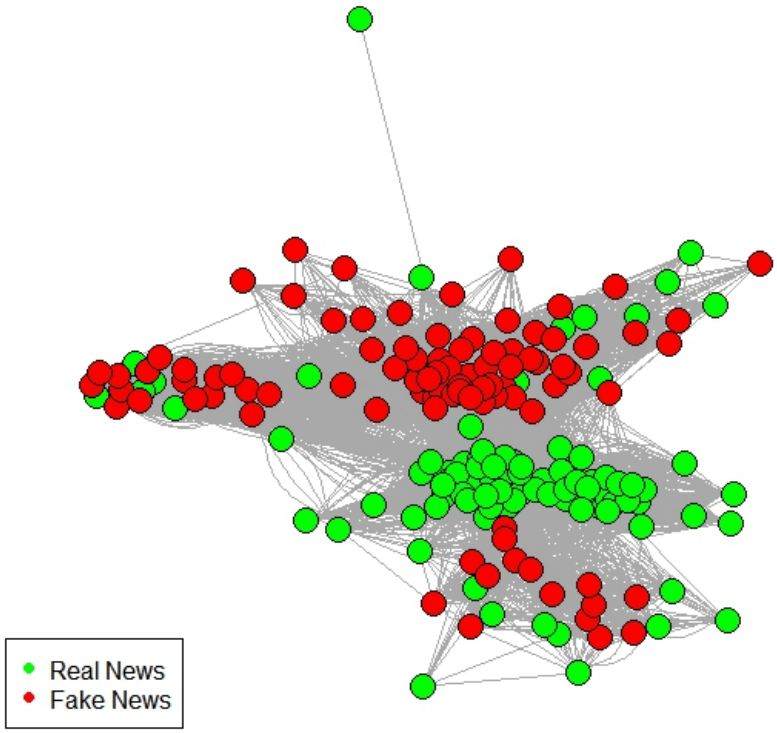

Figure 10: Grafo engagement gossip-news

Possiamo notare dai grafi (figure 9 e 10) come sia presente la tendenza dei nodi dello stesso colore di stare aggregati suggerendo la possibilità che la gran parte delle interazioni che gli utenti hanno rimangano all'interno della comunity corrispondente.

3.4 Affidabilità utenti

Nella sucessiva analisi ci siamo concentarti sugli utenti. A ciascun utente è stato attribuito un valore di affidabilità che descrive il rapporto tra le real-news con cui l'utente ha interagito e il numero totale di interazioni che ha fatto con le notizie. Quindi più un utente è affidabile più questo interagisce con un numero maggiore di real-news rispetto alle fake-news, viceversa, invece, un utente è definito meno affidabile se interagisce con un numero maggiore di fake news rispetto alle real news prese in studio.

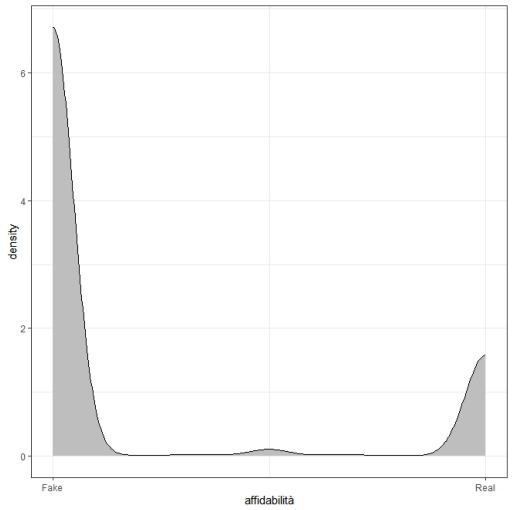

Figure 11: Affidabilità degli utenti political-news

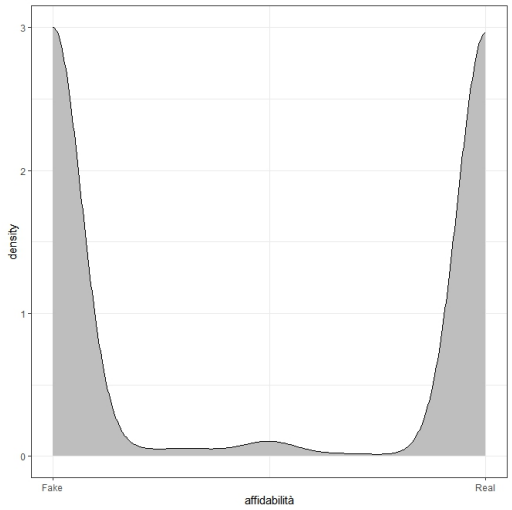

Figure 12: Affidabilità degli utenti gossip-news

Dalle figure 11 e 12 possiamo notare la distribuzione basata sull'affidabilità calcolata per ciacun utente, quindi nell'asse delle ascisse troviamo il valore di affidabilità mentre nell'asse delle ordinate la densità rispetto al numero di utenti. Nella figura 11 possiamo notare una notevole differenza nel numero di utenti affidabili e questo è sicuramente dovuto dalla differenza di utenti che hanno interagito con tali notizie, come visto nella figura 5.
In entrambe le figure possiamo notare un evidente polarizzazione degli utenti.

3.5 Rete dei follower

Nell'ultima analisi, abbiamo creato un grafo che illustra la rete di follower per verificare se le connessioni tra gli utenti possono essere influenzate dall'indice di affidabilità calolato in precedenza (capitolo 3.4). In questo grafo, ogni nodo simboleggia un utente che ha interagito con almeno una notizia, invece gli archi direzionali tra i nodi indicano le relazioni di follower (seguito) tra due soggetti. Il colore di ciascun nodo riflette invece il grado di affidabilità.

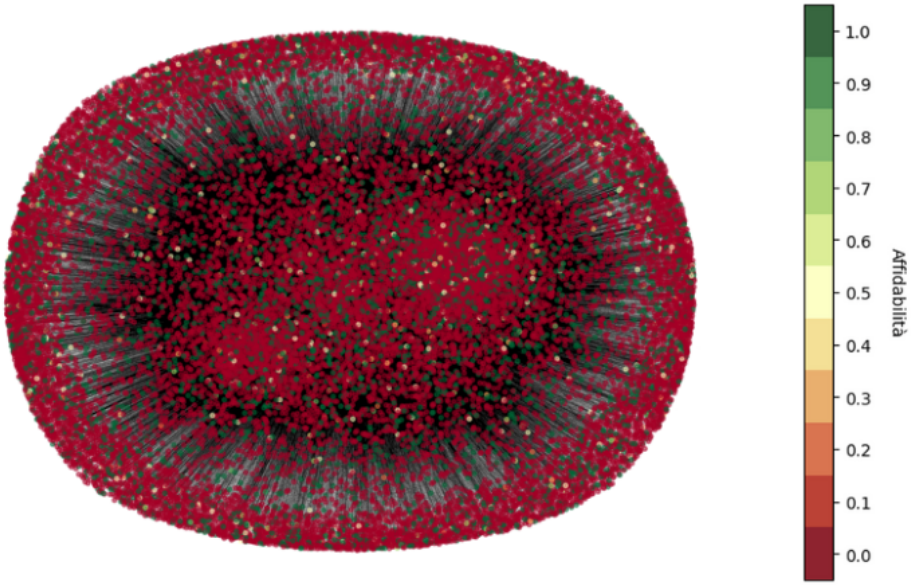

Figure 13: Rete dei follower utenti degli political-news

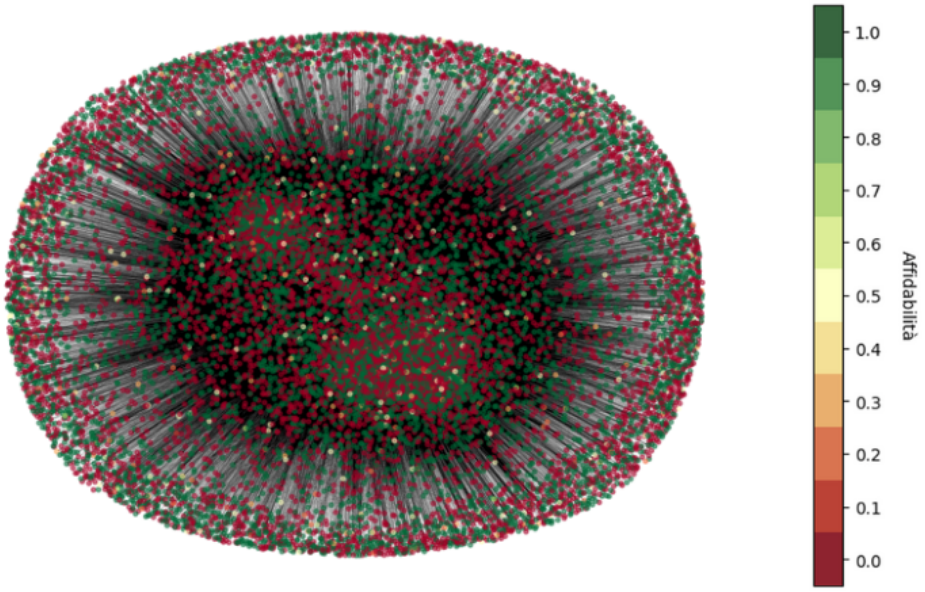

Figure 14: Rete dei follower utenti degli gossip-news

Osservando le figure 13 e 14, entrambe mostrano una vasta componente connessa centrale. Importante è notare come non sia presente una chiara correlazione tra le connessioni degli utenti e il loro grado di affidabilità, mostrando quindi come queste non si influenzino nel nostro caso di studio.

4 Risultati e conclusioni

Prima di trarre le conclusione della nostra ricerca è necessario analizzare i possibili *Biases* presenti nei nostri dati. Il pricipale è sicuramente la selezione delle notizie prese come caso di studio. É stato necessario filtrare le notizie per analizzarne solo un numero ridutto, nel nostro caso abbiamo optato per selezionare le medesime notizie utilizzate nell' articolo "Exploiting Tri-Relationship for Fake News Detection" di Shu, Kai and Wang, Suhang and Liu e Huan[5]. Un'ulteriore bias che ne deriva è sicuramente di natura temporale poichè la selezione delle notizie appertiene ad un periodo prolungato. Infine l'ultimo è associato a come vengono salvati i tweet degli utenti poichè vengono salvati solo quelli in cui è presente una citazione all'articolo corrispondente, quindi sono stati escluse tutte quelle possibili interazione che potevano comunque essere connesse all'articolo ma che non presentavano un'associazione diretta.

Prendendo quindi in esame il dataset utilizzato possiamo affermare che nel nostro caso di studio l'engamente delle fake news è maggiore rispetto alle notizie vere, inoltre abbiamo riscontrato come le interazioni di utenti comuni tra le notizie portino quest'ultime ad avere la tendenza ad essere raggruppate in cumunity basate sulla veridicità degli articoli. Abbiamo pututo notare anche l'estrema polarizzazione che gli utenti tendono ad avere rispetto alle loro interazioni tra real e fake news.

4.1 Futuri sviluppi

Ci sono molteplici applicazioni potenziali che possono essere derivate dall'utilizzo di questo dataset e del documento correlato. Una delle più significative è lo studio approfondito dei vari problemi associati alla diffusione delle fake news sui social media. Questo include la rilevazione delle fake news, l'analisi della loro evoluzione nel tempo e la ricerca di strategie efficaci per mitigare il loro impatto.
Inoltre, il dataset offre la possibilità di essere ampliato con l'aggiunta di nuove notizie. Questo permetterebbe di studiare l'evoluzione continua e dinamica della diffusione delle fake news sui social media, in questo modo, sarebbe possibile analizzare in modo più dettagliato come queste notizie false si propagano nel corso del tempo. Inoltre, potremmo osservare come i temi legati alle fake news possano cambiare e adattarsi in risposta a vari fattori. Questo tipo di analisi potrebbe fornire intuizioni preziose per comprendere e contrastare la diffusione delle fake news.

References

[1] David G. Rand Gordon Pennycook. The psychology of fake news. *LIFE MEDICAL SCIENCES JOURNALS*, 2021.

[2] David Lanius Romy Jaster. *Che cosa sono le Fake News?* Società editrice il Mulino, 2018.

[3] Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. Fakenewsnet: A data repository with news content, social context and dynamic information for studying fake news on social media. *arXiv preprint arXiv:1809.01286*, 2018.

[4] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explorations Newsletter*, 19(1):22–36, 2017.

[5] Kai Shu, Suhang Wang, and Huan Liu. Exploiting tri-relationship for fake news detection. *arXiv preprint arXiv:1712.07709*, 2017.

[6] Wikipedia. Elezioni parlamentari negli stati uniti d'america del 2018, 2024.