Santin Simone 886116

Corso di Sistemi Complessi Modelli e Simulazione:

Simulazione di contromisure alla diffusione di disinformazione su un social network

Sommario

1 Introduzione

La diffusione di disinformazione (o *fake news*) sui social network, a partire dalla pandemia di Covid-19, è diventata sempre più rilevante, e alcune piattaforme faticano a contrastare efficacemente il fenomeno, l'obiettivo di questo progetto è quindi testare in un ambiente simulato ad agenti come eventuali contromisure possano fermare la diffusione di queste fake news.

Per la realizzazione di questo esperimento è stato preso come riferimento il paper del Dipartimento di Informatica dell'Imperial College di Londra intitolato *"Can We Stop Fake News? Using Agent-Based Modelling to Evaluate Countermeasures for Misinformation on Social Media"*, in particolare, la struttura dell'ambiente di simulazione, dei singoli agenti e di alcune contromisure implementate si ispira direttamente a tale studio.

Più nel dettaglio, ad ogni step temporale, ciascun agente:

- genera una notizia in funzione del proprio livello di ingenuità;
- decide se condividere una notizia ricevuta;
- eventualmente segnala la notizia, nel caso scelga di non condividerla.

L'obiettivo principale dell'esperimento, considerando i tre tipi di agenti — *Non-believer*, *Gullible* e *Susceptible* (che inizialmente rappresentano la maggioranza) — è osservare, con e senza l'applicazione delle contromisure, come gli agenti *Non-believer* e *Gullible* influenzano i *Susceptible*, "convertendoli" tramite la diffusione di notizie vere o false. In particolare, si vuole analizzare l'efficacia delle contromisure adottate nel limitare la propagazione delle *fake news* e favorire la diffusione delle notizie veritiere, permettendo così una maggior parte di conversioni di agenti *Susceptible* in agenti *Non-believer*.

Il lavoro ha richiesto un lungo periodo di calibrazione, al fine di ottenere una simulazione il più possibile realistica e bilanciata, evitando un modello eccessivamente efficace nel

bloccare le *fake news* e cercando di riprodurre dinamiche plausibili osservabili nella realtà

2 Social Network come sistema complesso

I social network rappresentano un tipico esempio di sistema complesso, in cui il comportamento globale emerge dalle interazioni tra milioni di utenti. Ogni utente può scegliere se pubblicare contenuti, condividere notizie o segnalare informazioni false. Queste decisioni, pur essendo prese a livello individuale, sono fortemente influenzate dal comportamento degli altri utenti e, a loro volta, contribuiscono a modificarlo.

La diffusione delle informazioni, in particolare delle fake news, dipende non solo dalle singole scelte, ma anche dalla struttura della rete: utenti con un alto numero di connessioni possono amplificare rapidamente la portata di una notizia, raggiungendo vaste porzioni della rete. Piccoli eventi locali possono così innescare fenomeni di viralità su larga scala, difficili da contenere.

Il comportamento complessivo della rete sociale non può essere compreso analizzando solo i singoli utenti isolati, ma emerge dall'insieme delle interazioni e dall'equilibrio dinamico tra la diffusione delle informazioni e i meccanismi di controllo e moderazione.

In questo progetto si è sviluppato un modello che cerca di riprodurre il funzionamento dei social network reali. Ogni agente, oltre a svolgere le proprie azioni individuali, è influenzato dal comportamento degli agenti con cui è collegato nella rete.

Gli agenti modellati seguono l'approccio dei *model-based reflex agents* descritto da Russell e Norvig. Ciascun agente mantiene uno stato interno costituito dal proprio livello di credulità, dalla memoria delle notizie già ricevute e dalle penalità accumulate in caso di segnalazioni errate. Le azioni, come la condivisione o il reporting, dipendono non solo dalle percezioni attuali, ma anche dallo storico delle proprie interazioni precedenti e dai segnali di moderazione presenti nel sistema.

3 Descrizione del Modello

Lo sviluppo di questo esperimento ruota attorno a un meccanismo di creazione, condivisione e segnalazione delle notizie. Poiché risulta impossibile, in un sistema ad agenti, replicare fedelmente il contesto e i molteplici fattori che distinguono una *fake news* da una notizia vera, è stato introdotto un *credibility score* per ogni news, lo *score* infatti rappresenta l'elemento su cui si basa la capacità degli agenti di distinguere una notizia vera da una falsa.

3.1 Contromisure implementate

Le contromisure implementate nel sistema si basano su un meccanismo di segnalazione, tramite il quale un agente può segnalare una notizia quando la riceve, e segnalare anche il relativo agente che gliel'ha inoltrata.

Le contromisure applicate possono quindi essere raggruppate in tre categorie in base alle segnalazioni.

3.1.1 Rimozione delle news

Ogni volta che una notizia riceve una segnalazione, il suo contatore di *report* viene incrementato. Se il numero di segnalazioni raggiunge o supera quota tre, la notizia viene contrassegnata come sospetta e, ogni volta che un utente la visualizza, viene mostrato un avviso che indica la possibilità che si tratti di una *fake news*. Contestualmente, lo *score* di credibilità della notizia viene penalizzato in misura pari al numero di segnalazioni accumulate. Quando la notizia raggiunge cinque segnalazioni, viene automaticamente rimossa dal sistema.

3.1.2 Rimozione degli agenti

Contestualmente alla gestione delle notizie, anche gli agenti possono essere soggetti a segnalazioni. Ogni volta che un agente riceve una segnalazione, il suo contatore di *report* viene incrementato. Quando il numero di segnalazioni raggiunge o supera quota tre, tutte

le notizie da lui create nei successivi step verranno automaticamente contrassegnate come sospette. Al raggiungimento di cinque segnalazioni, analogamente a quanto avviene per le notizie, l'agente viene rimosso dal sistema.

Inoltre, prevenire eventuali abusi del sistema di *report* da parte di agenti malintenzionati, sono previste delle penalizzazioni in caso di segnalazioni errate. In particolare, quando un agente segnala come falsa una notizia che in realtà è vera, e di conseguenza anche il suo autore, viene temporaneamente sospesa la sua possibilità di effettuare nuove segnalazioni per due step. Dopo cinque segnalazioni errate, l'agente viene definitivamente rimosso dal sistema.

3.1.3 Riduzione visibilità agente

Come ultima contromisura è stato implementato quello che, nel gergo tecnico, viene definito *shadow banning*. In pratica, per ogni segnalazione ricevuta, viene ridotta la visibilità dell'agente, ossia diminuisce il numero di agenti vicini a cui vengono inoltrate le notizie da lui condivise. La riduzione è progressiva: ad ogni segnalazione la visibilità si abbassa del 20%, fino a raggiungere il 100% dopo cinque segnalazioni, corrispondente alla rimozione completa, ovvero al *ban* dell'agente.

3.2 Struttura delle news

Le news sono oggetti realmente implementati nel sistema. Ogni news è dotata di:

- un ID univoco,
- una lista di agenti che l'hanno condivisa,
- un flag che indica se è stata segnalata come sospetta,
- il numero totale di report ricevuti,
- il *credibility score*.

Per quanto riguarda il calcolo del *credibility score*, questo viene aggiornato a ogni condivisione della notizia, ad eccezione del momento della sua creazione. In questo modo si garantisce un sistema più bilanciato, evitando che i *non-believer* possano riconoscere troppo facilmente una fake news.

Al momento della creazione, il *credibility score* viene inizializzato a 0.5 per tutte le notizie e successivamente aggiornato come segue:

$$probGullible = (botShares * 0.0 + gullibleShares * 0.2 + susceptibleShares * 0.4 + nonBelieverShares * 0.8\ ) - newsReports * 0.05$$

Questo meccanismo consente di simulare in modo più realistico il comportamento di un utente di fronte a una notizia.

3.3 Struttura degli agenti

In questo esperimento, gli agenti si distinguono per due caratteristiche principali. La prima è la categoria di appartenenza, che determina il numero delle loro connessioni:

- **Influencer:** 10 connessioni con gli altri agenti
- **User:** 5 connessioni con gli altri agenti
- **Bot:** 3 connessioni con gli altri agenti

Per rendere il sistema piu dinamico, ogni 3 step è stata aggiunta una probabilità del 50% di creare nuove connessioni tra due agenti per ogni tipo di agente.

La seconda, e più importante ai fini dell'esperimento, è il loro livello di ingenuità, ossia *non believer, gullible*, e *susceptible*, di seguito è spiegata la loro struttura.

3.3.1 Agenti Non-Believer

Gli agenti *non-believer* sono quelli che non credono alle *fake news*. Ad ogni step hanno una probabilità del 50% di creare e inoltrare una notizia vera, questa probabilità è stata inserita in modo da rallentare il processo di testing. Quando ricevono una notizia poi, basandosi sul suo *credibility score* e se la notizia è contrassegnata come sospetta, decidono se inoltrarla oppure segnalarla, piu precisamente:

- Notizia sospetta AND Score < 0.5 : la notizia viene segnalata
- Score ≥ 0.7 : la notizia viene condivisa
- Score ≤ 0.3 : la notizia viene segnalata
- 0.3 < score < 0.7: 20% di probabilità di condividere la notizia

All'inizio della simulazione, i *non-believer* costituiscono il 20% degli agenti presenti nel sistema.

3.3.2 Agenti Gullible

Al contrario dei *non-believer*, l'obiettivo degli agenti *gullible* è rendere virali le *fake news*. Ad ogni step, anche loro hanno la stessa probabilità (50%) dei *non-believer* di creare una notizia, ma in questo caso si tratta di una *fake news*. Quando ricevono una notizia, la condividono consapevolmente se si tratta di una *fake news*. Se invece la notizia è vera, tentano di ostacolarne la diffusione, ossia: con una probabilità del 80% la segnalano, altrimenti la inoltrano, in questo modo è possibile garantire un maggior realismo al modello. Allo stato iniziale, così come per gli agenti *non-believer*, i *gullible* costituiscono il 20% del sistema, in modo da creare un modello bilanciato.

3.3.3 Agenti Susceptible

Gli agenti *susceptible* sono al centro dell'esperimento, in quanto rappresentano l'indicatore principale dell'andamento della simulazione e dell'efficacia delle contromisure applicate.

Questi agenti non creano notizie durante il proprio turno, ma, ogni volta che ne ricevono una, aggiornano i propri contatori di esposizione alle notizie vere e a quelle false. Quando ricevono una nuova notizia, confrontano il livello di esposizione accumulato: se l'esposizione alle notizie vere è superiore a quella alle notizie false, hanno una certa probabilità di convertirsi in *non-believer*, probabilità che cresce in funzione del numero di agenti *non-believer* vicini, piu precisamente:

$$probNonBeliever = 0.05 * nonBelieverNeighbor$$

Al contrario, se l'esposizione alle *fake news* prevale, hanno una probabilità di diventare *gullible*, influenzata dal numero di vicini *gullible*.

$$probGullible = 0.05 * gullibleNeighbor$$

Inoltre, se la notizia ricevuta è contrassegnata come sospetta e presenta un *credibility score* molto basso, l'agente avrà una probabilità aumentata di convertirsi in *non-believer*.

$$probNonBeliever = 0.1 * nonBelieverNeighbor$$

All'avvio della simulazione, i *susceptible* costituiscono la maggioranza del sistema, rappresentando il 60% degli agenti.

3.3.4 Agenti Bot

I *bot*, proprio come accade nei social network reali, hanno un unico compito: generare e diffondere *fake news* all'interno del sistema. Ad ogni step, infatti, creano e condividono una nuova *fake news*, contribuendo così alla propagazione della disinformazione. Questo comportamento riproduce fenomeni realmente osservati, come ad esempio la diffusione massiccia di notizie false da parte di bot automatizzati prima di eventi rilevanti, quali elezioni politiche o situazioni di crisi sanitaria.

4 Risultati Simulazione

Per testare il modello, la simulazione è stata eseguita per 100 step, inizializzando il sistema con 100 utenti normali, 100 *influencer* e 5 *bot*. I risultati ottenuti, presentati di seguito, si sono dimostrati stabili nella maggior parte delle esecuzioni, con variazioni minime tra una simulazione e l'altra. Anche aumentando la scala del sistema — portando a 1000 gli utenti normali, a 100 gli *influencer* e a 50 i *bot* — i risultati si sono mantenuti coerenti. Tuttavia, per ragioni di rapidità e ottimizzazione dei tempi di simulazione, si è scelto di condurre i test principalmente con un numero ridotto di agenti.

4.1 Scenario: Nessuna contromisura

Come primo scenario è stato analizzato il sistema senza alcuna contromisura, ossia rimuovendo completamente le opzioni di segnalazione, e i risultati ottenuti sono i seguenti:

Figura 1: Agents' credulity senza contromisure

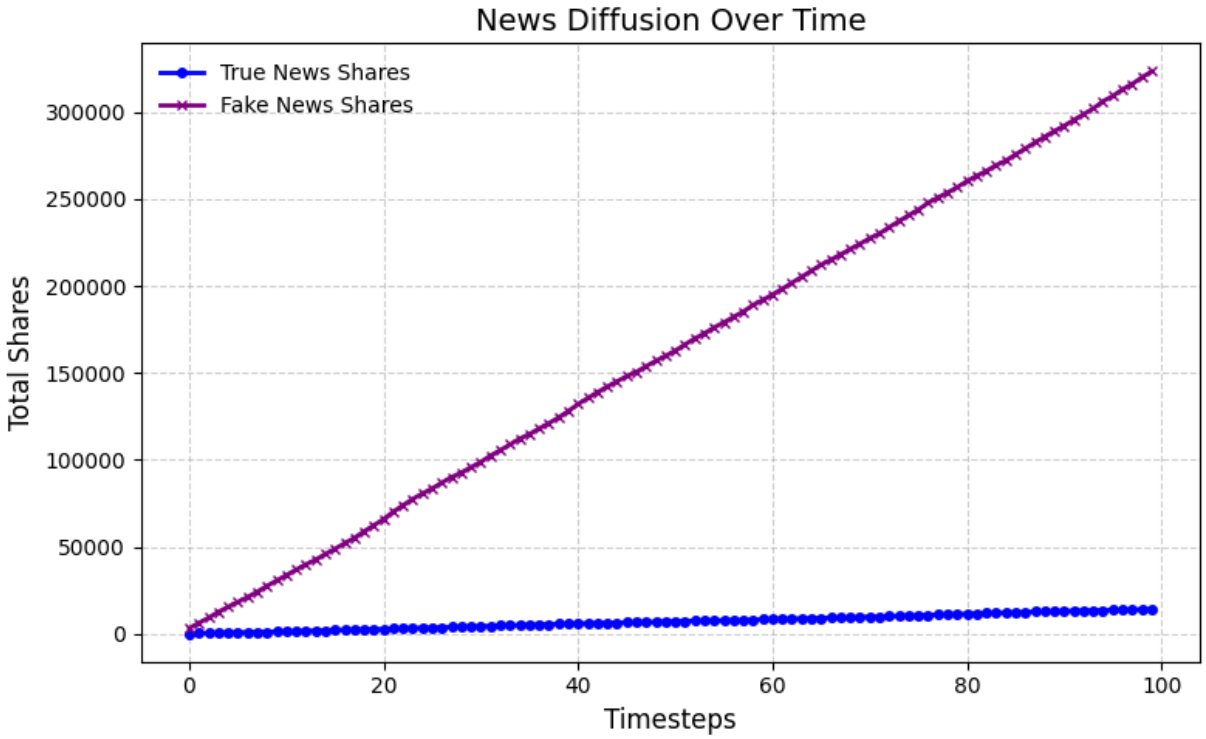

Figura 2: diffusione delle notizie senza contromisure

Come si può osservare, quasi tutti gli agenti *susceptible* diventano rapidamente *gullible*, mentre gli agenti *non-believer* mostrano un andamento abbastanza stabile. Per quanto riguarda la condivisione delle notizie, si nota che le notizie *fake*, al crescere degli step, e quindi con l'aumento del numero di agenti *gullible*, aumentano in modo esponenziale, mentre le notizie *true* rimangono relativamente stabili. Questo scenario potrebbe sembrare poco realistico; tuttavia, se si considera il comportamento di alcuni social network privi di controlli, come determinati forum online in cui proliferano teorie complottiste di ogni tipo, non appare così distante dalla realtà.

4.2 Scenario: Rimozione degli utenti

Successivamente, come prima contromisura, è stata applicata la rimozione degli utenti, testando il numero di report ricevuti dagli agenti e includendo anche un controllo sugli eventuali report errati.

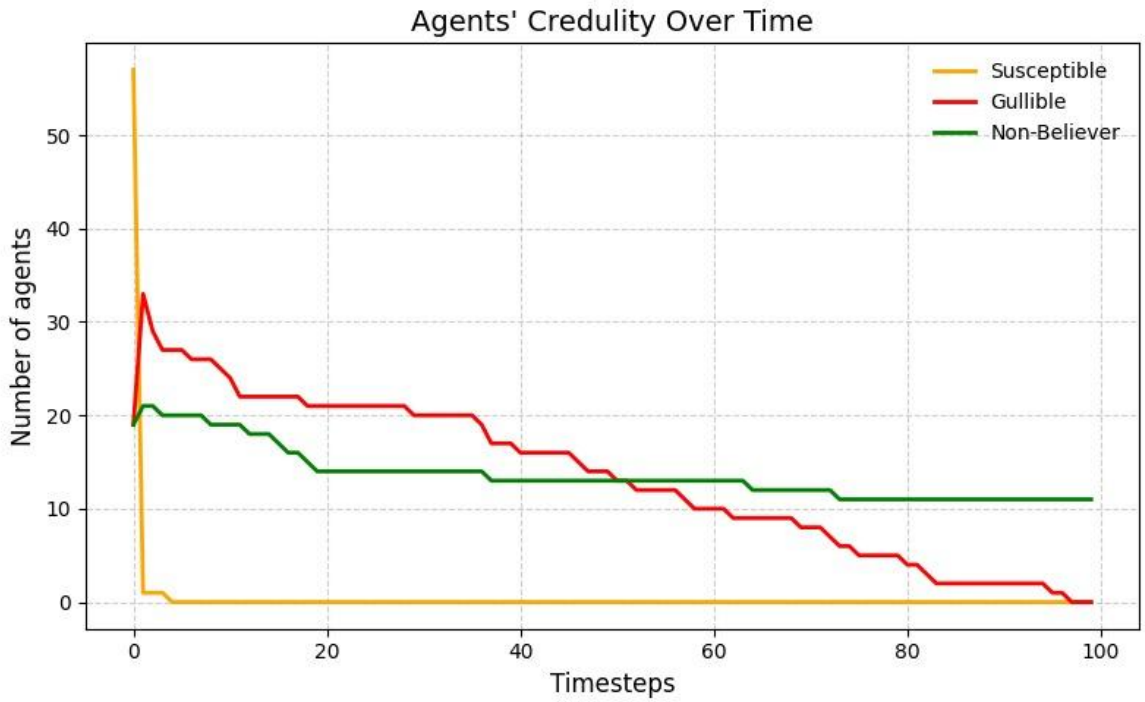

Figura 3: agents'credulity con report sugli user

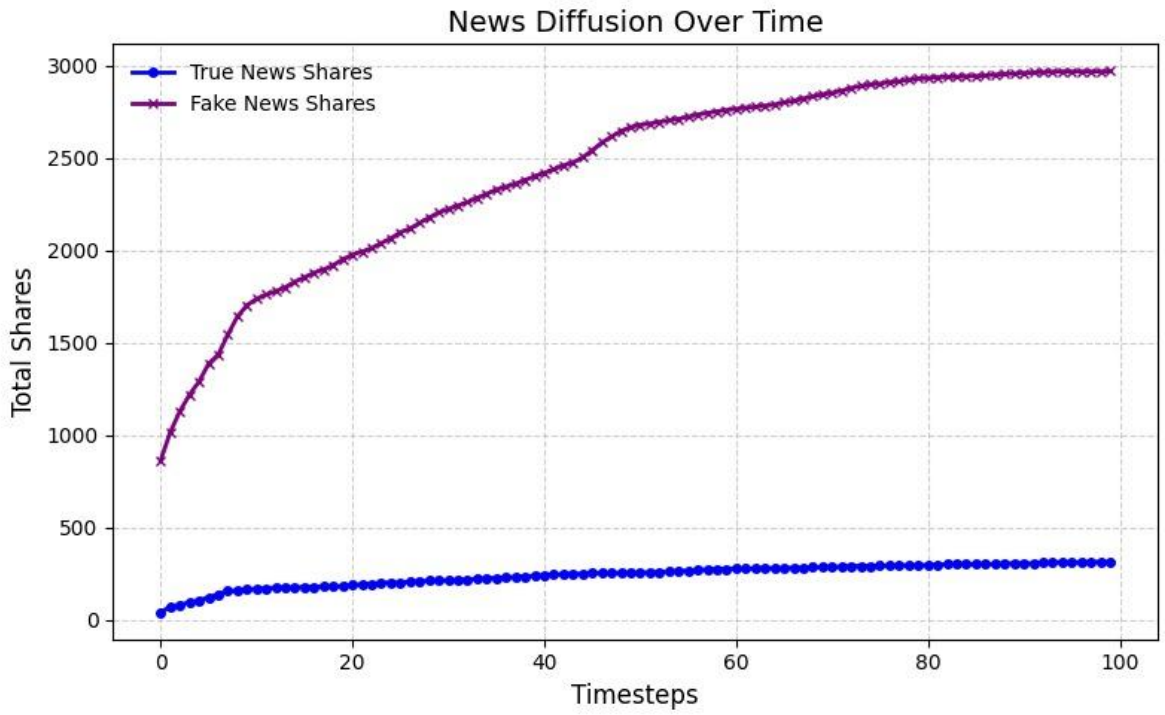

Figura 4: diffusione delle news con il report degli user

Attivando il sistema di *report* da parte degli *user*, si osserva un drastico calo degli agenti *gullible* rispetto allo scenario precedente privo di contromisure. Tuttavia, tale riduzione non è sufficiente a mantenere il fenomeno sotto controllo: nei primi step il numero di *gullible* supera ancora quello dei *non-believer*, per poi scendere al di sotto di esso negli step successivi, evidenziando il progressivo effetto della logica di *ban* col passare del tempo ma dimostrando come la mancanza di controlli sulle notizie fake, porti a una generazione maggiore di agenti *gullible* rispetto a *non -believer*.

Per quanto riguarda invece il grafico relativo alla condivisione delle notizie, si osserva che, in assenza di controlli diretti sui contenuti, il numero di *fake news* condivise cresce rapidamente, superando di molto quello delle notizie *true*, che, pur registrando un lieve incremento, rimangono relativamente stabili.

4.3 Scenario: Rimozione delle news

Per quanto riguarda la rimozione delle news, è stato testato il sistema di *report*, includendo sia la segnalazione delle notizie sospette sia il meccanismo di rimozione delle

stesse, i risultati ottenuti sono quindi i seguenti:

Figura 5: Agents'credulity con rimozione delle news

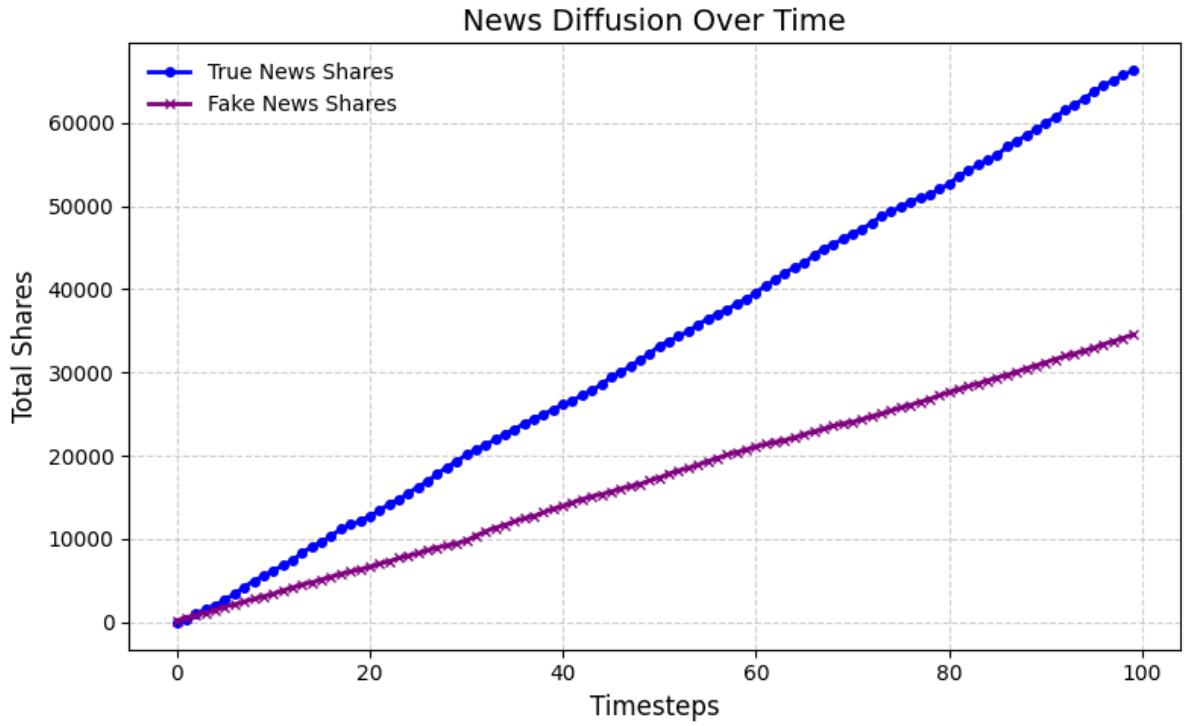

Figura 6: diffusione delle news con rimozione di esse

In questo scenario si osserva che, ostacolando la diffusione delle *fake news*, vi è una rapida crescita degli agenti *non-believer*: infatti, limitando l'esposizione dei *susceptible* alle notizie false, aumenta la probabilità che essi diventino *non-believer*. Tuttavia, poiché non viene gestita parallelamente la rimozione degli agenti, il numero di *gullible* rimane comunque elevato.

Per quanto riguarda la diffusione delle notizie, si nota che, pur con un maggiore controllo sui contenuti, il numero di *fake news* condivise rimane ancora alto. A differenza però dello scenario con la rimozione degli *user*, il numero di *true news* condivise cresce progressivamente, superando quello delle *fake news* con l'aumentare degli step. Questo conferma l'efficacia delle contromisure adottate, che riescono a ostacolare correttamente la diffusione delle *fake news*.

4.4 Scenario: contromisure complete

Come ultimo scenario sono state applicate tutte le contromisure, ovvero il *report* degli utenti, il *report* delle notizie e lo *shadow banning*. I risultati ottenuti sono i seguenti:

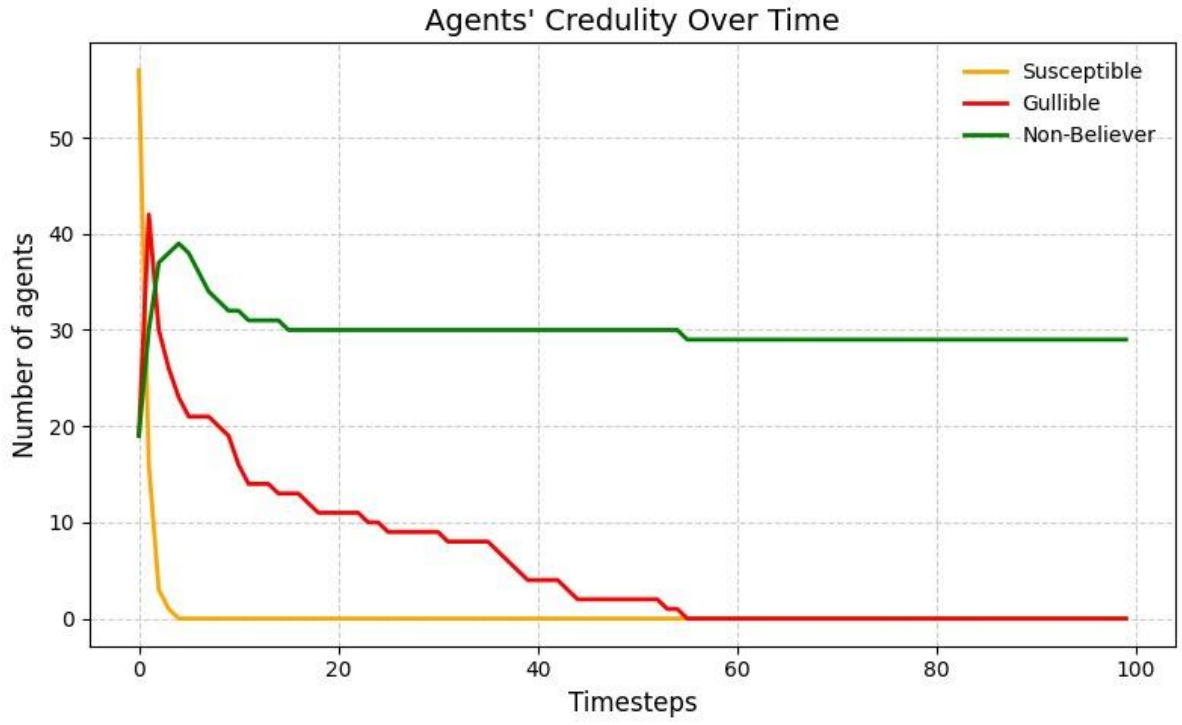

Figura 7: agents'credulity con le contromisure

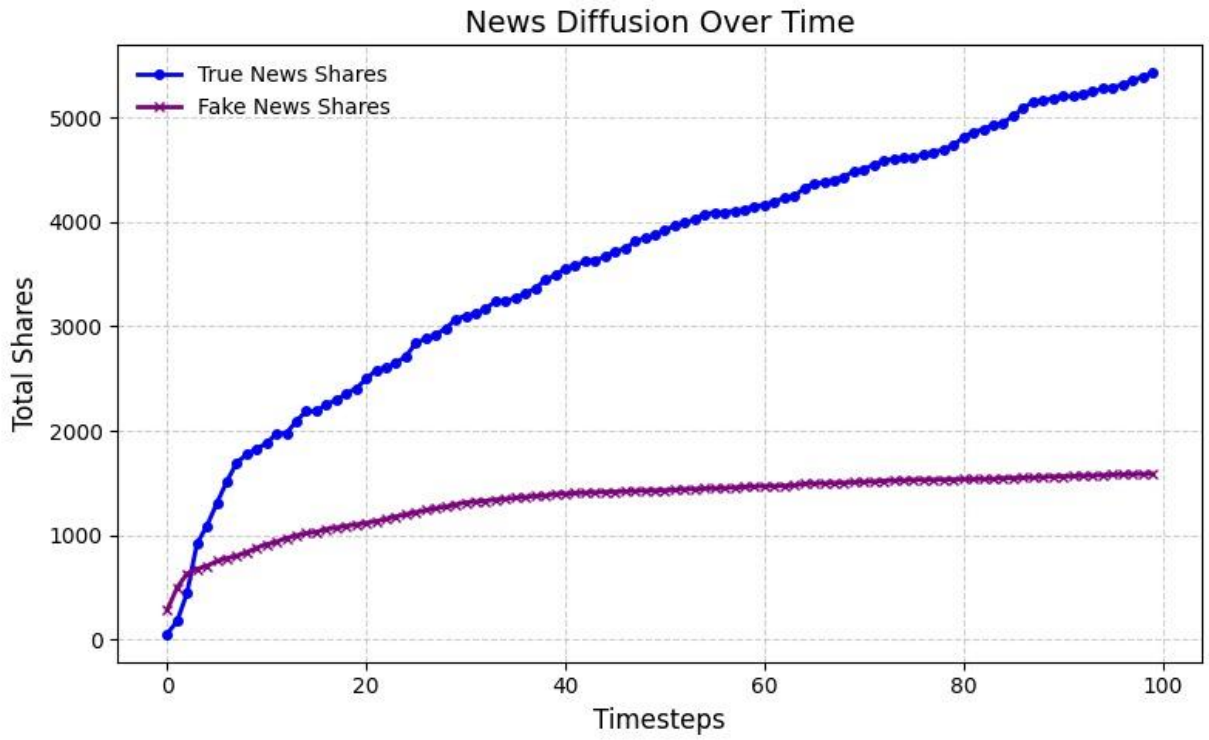

Figura 8: diffusione delle news con contromisure

Come si può osservare, applicando tutte le contromisure il sistema riesce a fermare efficacemente la diffusione delle *fake news*, con un netto aumento degli agenti *non-believer* e una drastica riduzione degli agenti *gullible* con il passare degli step.

Per quanto riguarda la diffusione delle notizie, si nota che, a differenza degli scenari precedenti, si verifica finalmente un'inversione nella distribuzione delle notizie: il sistema riesce ora a bloccare la diffusione delle *fake news*, favorendo una maggiore condivisione delle notizie *true*, dimostrando così che l'applicazione combinata di tutte le contromisure porta a risultati positivi.

4.5 Analisi contromisure

Dopo aver dimostrato l'applicazione delle contromisure, è possibile constatare che l'adozione combinata di diversi meccanismi di controllo consente di contrastare in modo efficace la diffusione delle *fake news*, agendo sia sugli agenti che sui contenuti e producendo effetti significativi sull'evoluzione della rete.

- **Rimozione degli utenti**: la moderazione basata sui *report* sugli agenti permette di ridurre progressivamente il numero di *gullible*, sebbene nelle prime fasi il numero di questi agenti rimanga comunque elevato. Questo meccanismo agisce più efficacemente nel lungo periodo, contribuendo a stabilizzare il sistema, tuttavia va a perdere sul controllo delle news, infatti il basso numero di agenti *gullible* verso la fine deriva da una conseguenza di molteplici ban, piu che conversioni corrette di susceptible.
- **Rimozione delle notizie**: il *reporting* dei contenuti consente di limitare l'esposizione dei *susceptible* alle *fake news*, favorendo così la transizione verso lo stato di *non-believer*. Tuttavia, se non accompagnata da un controllo sugli agenti, non riesce a ridurre drasticamente il numero complessivo di *gullible*.
- **Combinazione di tutte le contromisure**: l'integrazione simultanea dei meccanismi di moderazione sui contenuti e sugli agenti si dimostra la strategia più efficace. Si osserva una netta riduzione dei *gullible* grazie alla logica di rimozione, un forte incremento dei *non-believer* grazie alla logica di rimozione e segnalazione delle notizie, con una progressiva predominanza delle notizie vere rispetto alle *fake news*.

5 Conclusioni

Il progetto ha consentito di sviluppare e analizzare un modello agent-based per simulare la diffusione di fake news in un contesto ispirato al funzionamento dei social network. La definizione di categorie di agenti con differenti ruoli (influencer, user, bot) e livelli di suscettibilità cognitiva ha permesso di riprodurre dinamiche di propagazione realistiche, tipiche di sistemi complessi reali.

L'introduzione progressiva delle contromisure ha evidenziato come interventi mirati possano modificare significativamente l'evoluzione della rete e il comportamento complessivo degli agenti. In particolare:

- Il controllo sugli **agenti** (tramite meccanismi di reporting e ban) consente di ridurre la presenza di soggetti propensi alla diffusione incontrollata di contenuti falsi;
- Il controllo sui **contenuti** (tramite flagging e rimozione delle notizie sospette) agisce direttamente sull'esposizione informativa degli agenti, influenzando le transizioni cognitive tra gli stati di credulità;
- La combinazione sinergica di entrambe le strategie si è dimostrata la più efficace, portando a una progressiva prevalenza delle notizie veritiere e a un contenimento stabile della diffusione delle fake news.

Pur rappresentando una semplificazione rispetto alla complessità dei sistemi sociali reali, il modello sviluppato consente di analizzare in modo dettagliato i meccanismi di propagazione della disinformazione e di valutare l'efficacia di diverse politiche di moderazione.

In conclusione, il lavoro svolto conferma come l'approccio agent-based costituisca uno strumento flessibile e potente per l'analisi di fenomeni complessi quali la disinformazione online e per la progettazione di possibili strategie di contrasto efficaci.