Politecnico di Torino

Politecnico di Torino

Corso di Laurea
A.a. 2022/2023
Sessione di Laurea Dicembre 2023

Bilanciare espressione e sicurezza nella realtà virtuale sociale

Relatori:
Andrea Bottino
Francesco Strada

Candidati:
Caterina Nuovo

Sommario

Introduzione

Nel panorama delle crescenti innovazioni tecnologiche, l'evoluzione delle tecnologie di realtà aumentata e virtuale sta aprendo nuove prospettive per l'interazione e la condivisione di esperienze immersive. Queste tecnologie consentono agli utenti di immergersi in ambienti tridimensionali digitali, dove possono interagire con altre persone e oggetti virtuali in tempo reale. Questa innovazione sta rivoluzionando il modo in cui le persone si collegano tra loro e condividono informazioni, migliorando le esperienze sociali, stimolando l'espressione creativa e creando nuove opportunità per lo scambio di conoscenze su scala globale.

Le esperienze immersive multiutente possono servire a vari scopi, fungendo da strumenti di collaborazione professionale, ambienti di apprendimento, luoghi per incontri sociali ed eventi o piattaforme per giochi multiplayer. La loro versatilità li rende infatti adatti a soddisfare una vasta gamma di esigenze degli utenti.

Tra queste esperienze immersive emerge il concetto di Social VR, che rappresenta una fusione intrigante tra le piattaforme di social networking convenzionali e i videogiochi. In un ambiente di Social VR, gli utenti si trovano in spazi virtuali 3D dove sono immersi in attività coinvolgenti e possono interagire con altri utenti attraverso dispositivi di realtà virtuale (HMD). Questi servizi consentono agli utenti di

socializzare a uno a uno o in gruppi di varie dimensioni e di condividere esperienze con oggetti o ambienti virtuali.

Questo mix unico combina gli aspetti delle piattaforme di social networking, che permettono agli utenti di connettersi e interagire, con l'elemento dei videogiochi, offrendo esperienze immersive e interattive. Gli utenti possono quindi condividere esperienze simili a quanto fanno su piattaforme di social networking, ma all'interno di mondi virtuali in cui possono creare avatar, partecipare ad eventi e attività di gioco, o esplorare ambienti virtuali.

È rilevante notare che questi servizi possono funzionare in due modalità principali: da un lato, come canali per comunicazioni private tra individui, simili a conversazioni telefoniche o riunioni private in una sala conferenze; dall'altro lato, come un'interfaccia con una vasta quantità di utenti, la maggior parte dei quali sono sconosciuti tra loro. Quest'ultimo scenario consente di creare nuove comunità al di là delle distanze fisiche e offre nuovi canali per l'espressione creativa e l'innovazione, ma allo stesso tempo spesso vede alcuni utenti avere comportamenti inappropriati nei confronti degli altri, come abusi sessuali, bullismo, discorsi d'odio e molestie di vario tipo. Pertanto, è fondamentale che le piattaforme implementino misure di moderazione per prevenire o mitigare tali comportamenti indesiderati e assicurare un certo livello di benessere nelle comunità che frequentano questi spazi virtuali. È altresì importante considerare che gli utenti con

relazioni preesistenti potrebbero necessitare di strumenti di moderazione dei contenuti e di sicurezza differenti rispetto agli utenti a loro estranei. Un equilibrio delicato deve essere trovato tra garantire la sicurezza degli utenti e preservare la privacy e la libertà di espressione nelle interazioni virtuali. [1]

Stato dell’arte

Molestie nella social VR

La prima testimonianza pubblica di un episodio di molestia avvenuto in un ambiente virtuale risale al 2016, quando una donna di nome Jordan Belamire pubblicò un articolo intitolato “My First Virtual Reality Groping”, in cui descrisse come era stata molestata sessualmente da un altro giocatore mentre giocava a QuiVr. [2]. Col passare degli anni sono stati segnalati sempre più casi di abusi sessuali, bullismo, discorsi d'odio, razzismo, e varie forme di comportamenti inappropriati negli ambienti di social VR. Di seguito alcuni dati estrapolati da indagini che sono state svolte in ambienti di social VR per valutare la diffusione di episodi di abuso:

- In uno studio del 2017 sugli ambienti di social VR, 2 donne su 7 e 21 uomini su 99 hanno segnalato di aver subito molestie in social VR, mentre il 42% dei partecipanti ha dichiarato di aver assistito a episodi di abuso. [3]
- In una ricerca del 2018 che ha visto partecipare 600+ utenti della social VR, il 49% delle donne intervistate e il 38% degli uomini hanno dichiarato di aver subito molestie sessuali nella realtà virtuale. [4]
- In un’intervista sulle molestie negli ambienti di social VR fatta nel 2019, il 92% dei partecipanti ha riferito di aver assistito a episodi di molestie in social VR,

mentre il 72% ha dichiarato di aver sperimentato comportamenti che gli autori classificherebbero come molestie. [5]

- I ricercatori del The Centre for Countering Digital Hate (CCDH), dopo aver trascorso ore su VRChat, hanno stimato che gli utenti fossero esposti a comportamenti abusivi ogni sette minuti, tra cui minori esposti a contenuti sessuali grafici, bullismo, molestie sessuali e abusi nei confronti di altri utenti, compresi i minori, minori istruiti a ripetere insulti razzisti e punti di vista estremisti, minacce di violenza. [6]

Questo fenomeno ha imposto una necessità sempre più impellente di adottare misure rigorose di moderazione e sensibilizzazione per affrontarlo e creare un ambiente virtuale più sicuro e rispettoso per tutti gli utenti in cui questi possano godersi la loro esperienza.

Nel contesto degli ambienti di social VR, è possibile individuare tre principali tipi di molestie [5]. È importante notare che alcune di queste forme di molestia risultano esclusive degli ambienti virtuali a causa delle straordinarie possibilità e dei metodi di interazione avanzati offerti da tali ambienti, superando di gran lunga le capacità degli altri ambienti online.

- **Molestie verbali**: questo tipo di molestia coinvolge l'uso di parole offensive, insulti, minacce o commenti denigratori nei confronti di un utente. Come

vedremo tra poco, le molestie verbali nella social VR possono raggiungere un nuovo livello di intensità grazie alla comunicazione vocale diretta tra gli utenti. L'immersione nell'ambiente virtuale può accentuare il coinvolgimento emotivo delle vittime, rendendo le parole offensive ancora più impattanti.

- **Molestie fisiche**: le molestie fisiche in social VR si manifestano attraverso il contatto fisico indesiderato o non consensuale tra gli utenti, spesso utilizzando gli avatar o gesti virtuali. Questo tipo di molestia è specifico degli ambienti virtuali, a causa della possibilità di interagire fisicamente con gli avatar degli altri utenti, e può manifestarsi sotto forma di movimenti dirompenti, simulazione di abusi fisici, simulazioni di aggressioni sessuali, stalking virtuale e simulazioni di autolesionismo o suicidio. [7]
- **Molestie ambientali**: questo tipo di molestia coinvolge la manipolazione dell'ambiente virtuale per scopi molesti o offensivi. Gli utenti possono mostrare immagini inappropriate o disturbanti su schermi virtuali condivisi o lanciare oggetti virtuali per creare disagio o interrompere le interazioni degli altri utenti.

Questo lavoro di ricerca si concentrerà sull'esplorazione delle molestie verbali in particolare e delle pratiche di moderazione utilizzate per approcciare questo problema che, come risulta da vari studi e dal questionario preliminare che è stato condotto durante questo studio, è ancora molto presente negli ambienti di social

VR. Inoltre, il canale vocale è quello più utilizzato per la comunicazione in quanto anche più comodo rispetto a quello testuale a causa dei device di input.

Quali sono i fattori di rischio?

La social VR presenta elementi simili a quelli delle piattaforme multiutente più comuni, come piattaforme di videoconferenza, chat audio, giochi multiplayer e social media convenzionali, ma, come accennato prima, presentano anche caratteristiche uniche che comportano una serie di fattori di rischio distinti. [8] [9]

- **Anonimato e disinibizione online**: le persone mostrano comportamenti più estroversi, aggressivi o disinibiti online rispetto a quelli che mostrerebbero nella vita reale. Ciò può essere dovuto all'anonimato, all'assenza di conseguenze immediate per i propri comportamenti o alla distanza emotiva e fisica dalle persone coinvolte. Questa disinibizione può portare a comportamenti positivi, come l'apertura a nuove idee e alle relazioni sociali online, ma può anche portare a comportamenti negativi.
- **Potenzialità della realtà virtuale**: la spiccata componente immersiva della realtà virtuale, dovuta, ad esempio, alla comunicazione sincrona, all'utilizzo di avatar 3D che dà agli utenti la sensazione di vivere fisicamente ciò che il loro avatar vede e sente, o alle interazioni fisiche tra gli utenti, può far

percepire un episodio di abuso nel mondo virtuale come se fosse reale ed esacerbare così l'impatto e il danno sulle vittime.

- **Indizi sull'identità delle persone forniti dalla realtà virtuale**: ad esempio, la voce può fornire importanti indizi riguardo l'identità di un utente, come la sua nazionalità e il suo genere, e allo stesso tempo l'aspetto dell'avatar in alcuni casi potrebbe suggerire ulteriori informazioni, ad esempio l'altezza dell'avatar potrebbe dare indizi sull'età dell'utente. Questo rende inevitabilmente alcuni utenti più soggetti alle molestie di altri.
- **La natura effimera della realtà virtuale**: in questi ambienti gran parte delle interazioni avvengono in tempo reale e non vengono registrate, ad esempio voce e movimento. Questa caratteristica solleva nuove sfide nella moderazione simili a quelle che devono affrontare altre piattaforme di chat audio come Discord, dove le conversazioni possono svolgersi anche nei canali vocali senza lasciare tracce permanenti. [10]
- **La definizione di molestia è altamente soggettiva:** non esiste una definizione univoca di cosa è una molestia poiché è un concetto con delle forti basi sociali e culturali; inoltre, la maggior parte degli utenti si affida al loro buon senso e non legge politiche e tutorial. La percezione dei comportamenti dannosi si evolve nel tempo all'interno della società e quindi anche delle piattaforme di realtà virtuale: comportamenti precedentemente considerati innocui hanno iniziato ad essere visti, nel tempo, come dannosi,

prendiamo per esempio in considerazione quanto dibattito ha sollevato e sta sollevando il cosiddetto catcalling negli ultimi anni.

- **Mancanza di norme sociali standardizzate**: a differenza degli spazi fisici in cui le norme sociali sono spesso stabilite e regolate implicitamente, gli ambienti virtuali possono mancare di linee guida universalmente condivise e accettate. Questa assenza di un quadro normativo consolidato può creare un terreno fertile per comportamenti inappropriati, poiché gli utenti potrebbero trovarsi a navigare in un contesto sociale senza confini chiari.
- **Convivenza tra adulti e bambini all'interno degli stessi ambienti**: questo fattore presenta un ulteriore rischio di molestie, derivante dalle disparità nella concezione di ciò che è appropriato o inappropriato. La percezione differenziata tra le due fasce d'età può generare terreno fertile per abusi da entrambe le parti. Sebbene gli adulti possano esprimere preoccupazione per il fatto che i bambini interrompano e rovinino l'esperienza virtuale, i minori potrebbero anche subire comportamenti molesti da parte degli adulti. [11]

Quali sono le sfide?

- **Sviluppo di norme sociali adeguate**: questo processo è complicato dalla natura soggettiva della definizione di molestia. Ciò implica che ciò che può essere considerato inappropriato da un individuo potrebbe non esserlo per

un altro. Inoltre, la coesistenza di piattaforme indipendenti e culturalmente diverse aggiunge ulteriori complessità, poiché norme sociali differenti possono scontrarsi. È necessario trovare un equilibrio che rispetti la diversità culturale senza tollerare comportamenti dannosi.

- **Strumenti di prevenzione proattiva**: la prevenzione proattiva delle molestie richiede lo sviluppo di strumenti avanzati che siano in grado di identificare e prevenire situazioni di molestia prima che si verifichino. Questi strumenti dovrebbero essere in grado di riconoscere pattern comportamentali sospetti o linguaggio offensivo, contribuendo così a creare un ambiente virtuale più sicuro.
- **Moderazione scalabile**: un'altra sfida è garantire la scalabilità dei sistemi di moderazione per gestire grandi volumi di utenti e contenuti simultaneamente. Questo richiede l'implementazione di algoritmi avanzati e l'infrastruttura tecnologica necessaria per rispondere rapidamente alle segnalazioni di molestie e intervenire in modo efficace.
- **Approcci di moderazione per nuove forme di contenuto**: con l'evoluzione delle social VR, è essenziale sviluppare approcci di moderazione che vanno oltre la tradizionale analisi di testi, immagini e video. Questo include la moderazione di ambienti virtuali, comunicazione virtuale non verbale attraverso avatar e oggetti tridimensionali. Adattare gli strumenti di

moderazione a queste nuove forme di contenuto digitale rappresenta una sfida tecnologica significativa.

- **Termini di utilizzo e linee guida condivise**: l''elaborazione di termini di utilizzo e linee guida condivise tra diverse piattaforme rappresenta una sfida normativa. La collaborazione tra gli sviluppatori delle social VR è essenziale per stabilire standard uniformi che favoriscano un ambiente virtuale più sicuro e rispettoso.
- **Bilanciare sicurezza con privacy e libertà di espressione**: questa sfida consiste nel bilanciare efficacemente le misure di sicurezza, come il monitoraggio delle conversazioni o la limitazione dei gesti, con il rispetto della privacy degli utenti e la promozione della libertà di espressione nelle interazioni virtuali immersive. Trovare un equilibrio che protegga gli utenti senza compromettere l'esperienza e la libertà di comunicazione rappresenta una delicata sfida etica e tecnologica.

In sintesi, gli ambienti di social VR presentano sfide uniche legate alle molestie a causa delle loro possibilità e metodologie di interazione nuove e superiori rispetto agli altri ambienti online. Questo pone una responsabilità crescente sulle piattaforme per sviluppare strategie di moderazione efficaci al fine di creare esperienze virtuali sicure e rispettose per tutti gli utenti, in un contesto sempre più complesso di ambienti virtuali multiutente.

Moderazione

Le piattaforme online stabiliscono regole per moderare i contenuti, oltre quanto richiesto dalla legge, in base alla natura del servizio e alle aspettative degli utenti. Per alcuni contenuti illegali chiaramente definiti, come la pornografia infantile o i contenuti terroristici violenti, esiste consenso su come affrontarli. Tuttavia, per altri tipi di contenuti, come la disinformazione o l'incitamento all'odio, le linee guida sono meno chiare e le piattaforme devono decidere autonomamente come gestirle. Con l'evoluzione delle piattaforme digitali sono emerse politiche sempre più complesse per affrontare i contenuti dannosi e preservare la libertà di espressione degli utenti. Questi approcci di moderazione includono spesso linee guida della community, segnalazioni degli utenti e moderazione proattiva da parte sia di moderatori umani che di strumenti di apprendimento automatico, mentre le azioni possono variare dalla rimozione di contenuti illegali alle sanzioni per violazione delle policies della piattaforma.

Le pratiche di moderazione sono principalmente proattive o reattive:

- Le **pratiche proattive** hanno come obiettivo quello di evitare che gli utenti tengano comportamenti dannosi, ad esempio informando gli utenti delle regole, ponendo delle limitazioni alle azioni degli utenti per evitare disturbi o posizionando strategicamente i moderatori in modo da essere ben visibili.

- Le **pratiche reattive** vengono adottate in risposta a comportamenti dannosi e possono includere avvertimenti, silenziamento o rimozione degli utenti che violano le regole o causano danni.

Moderazione umana

La **moderazione umana** prevede che delle persone, chiamate moderatori, rivedano i contenuti per determinare se violano le regole della piattaforma.

- **Pro**: gli esseri umani possono comprendere il contesto, l'intento e le sfumature dei contenuti, prendendo decisioni più accurate in situazioni complesse.
- **Contro**: questo approccio è costoso in termini di tempo e risorse, quindi poco scalabile, questo può portare a mancanza di chiarezza nelle regole e nelle decisioni prese o ritardi nella moderazione dovuti alla disponibilità limitata di moderatori. Possono verificarsi errori umani e le opinioni personali dei moderatori possono influenzare le decisioni. I moderatori potrebbero esercitare abuso di potere. I moderatori si ritrovano spesso a confrontarsi con contenuti disturbanti, come violenza estrema e abusi, e a dover prendere decisioni difficili. [12]

Moderazione basata sulla comunità

La **moderazione basata sulla comunità** coinvolge gli utenti stessi nella segnalazione e nella valutazione dei contenuti.

- Pro: coinvolge gli utenti stessi nel reporting e nella valutazione dei contenuti, distribuendo il carico di lavoro. Riflette le norme della comunità.
- Contro: può portare a segnalazioni abusive, per esempio nel caso di possibilità di creare dei sondaggi per cacciare un utente da uno spazio condiviso, o norme culturali limitate poiché le norme all'interno di una comunità online possono basarsi su un insieme specifico di valori o credenze culturali che potrebbero non essere inclusivi o riflettere le opinioni di altre culture o gruppi di persone.

Moderazione automatizzata

Gli strumenti di moderazione automatizzata entrano in gioco quando i problemi legati alle dimensioni di una comunità online rendono impossibile la moderazione manuale e sono diventati una pratica consolidata nelle più tradizionali piattaforme online. Questi sono in grado di identificare i contenuti generati dagli utenti che violano le regole delle piattaforme utilizzando tecniche di matching o previsioni e vengono utilizzati dalle piattaforme per monitorare contenuti di vario genere, tra cui terrorismo, violazione della proprietà intellettuale, violenza grafica, toxic speech

(discorsi d'odio, insulti, minacce e bullismo), contenuti di natura sessuale, abusi su minori e rilevamento di spam o account falsi. Nella maggior parte dei casi, entrambe le tecniche vengono utilizzate nella moderazione dei contenuti online, con l'obiettivo di identificare, rimuovere o segnalare contenuti inappropriati o pericolosi, con l'obiettivo di ridurre la necessità di un intervento umano. [13]

Le **tecniche di matching** mirano a identificare se due contenuti sono uguali. Nel caso di contenuti testuali, si confrontano le parole con quelle presenti in delle blacklist definite dalle piattaforme. Nel caso di contenuti multimediali, come le immagini, si utilizzano funzioni di hash che trasformano il contenuto in un hash unico che viene poi confrontato con quello di contenuti problematici noti di cui viene tenuta traccia in database accessibili pubblicamente. Questo è possibile grazie all'esistenza di iniziative come quella del Global Internet Forum to Counter Terrorism [14], fondato da Facebook, Microsoft, Twitter e YouTube, che mantiene un database di contenuti estremisti, o il database PhotoDNA [15] di Microsoft, che cataloga materiale noto sullo sfruttamento minorile.

Le **tecniche di classificazione**, invece, tentano di assegnarlo ad una determinata categoria. La classificazione utilizza spesso l'apprendimento automatico, sfruttando modelli addestrati su dati etichettati dagli esseri umani. Questi modelli possono fare affidamento su funzionalità codificate manualmente o utilizzare

tecniche di apprendimento automatico per analizzare il testo e rilevare contenuti offensivi o dannosi.

Riportiamo ora una serie di considerazioni sulle sfide introdotte dai sistemi di moderazione automatizzati, ma applicate al caso specifico della toxic speech, che è quello su cui ci si concentra questo lavoro di ricerca; diverse aziende, tra cui Google, Twitter e Facebook, hanno sviluppato sistemi di classificazione per riconoscere la toxic speech.

Innanzitutto, gli algoritmi di intelligenza artificiale devono essere addestrati su grandi quantità di testo etichettato manualmente dall'uomo in base al livello di tossicità, al fine di creare sistemi di classificazione automatica in grado di individuare i commenti tossici, e gli esseri umani che eseguono l'annotazione dei dati vengono a loro volta esposti a contenuti disturbanti, che possono avere un impatto sulla loro salute mentale.

In secondo luogo, è necessaria una supervisione umana continua per verificare le prestazioni degli algoritmi e aggiornare i dati di addestramento per tenere conto delle nuove politiche e delle nuove forme di contenuti problematici.

In terzo luogo, quando la precisione degli algoritmi non è sufficiente, i moderatori umani sono chiamati a intervenire per risolvere le situazioni più complesse. Infatti, nonostante i progressi nell'automazione, gli attuali algoritmi spesso non riescono a

raggiungere i livelli di precisione richiesti a causa della complessità e dalla variabilità della lingua e possono produrre risultati inadeguati classificando erroneamente commenti innocenti come tossici o viceversa e portando potenzialmente ad un'eccessiva rimozione di contenuti e ad una limitazione della libertà di espressione degli utenti che frequentano le piattaforme. Anche le sfumature culturali e linguistiche, come nel caso della moderazione umana, possono aggiungere complessità alle pratiche di moderazione. La sfida principale in questo caso è comprendere il contesto e interpretare correttamente le parole, considerando che alcune parole possono essere utilizzate in modi diversi a seconda del contesto. Un esempio emblematico è uno studio del 2020 sui rischi della moderazione basata sull'utilizzo di IA, in cui sono state analizzate le prestazioni di un'API per la classificazione dei contenuti tossici online, Perspective, sviluppata da Jigsaw, per misurare i livelli di "tossicità" di una serie di post su Twitter. Lo studio ha confrontato i livelli di tossicità dei tweet di famose drag queen negli Stati Uniti con quelli di altri importanti utenti, in particolare dei nazionalisti bianchi, e ha rilevato che Perspective considera molti resoconti delle drag queen più tossici di quelli dei nazionalisti bianchi e di Donald Trump non riuscendo a considerare il contesto sociale quando misura i livelli di tossicità e a riconoscere i casi in cui le parole che potrebbero essere viste come offensive assumono significati diversi nel linguaggio LGBTQ o più in generale i casi in cui viene utilizzata la cosiddetta "mock impoliteness", che comprende una vasta gamma di interazioni, quali scherzi, prese

in giro, insulti scherzosi e umorismo. Lo studio suggerisce così che l'intelligenza artificiale potrebbe rafforzare i pregiudizi contro la comunità LGBTQ e limitarne la libertà d'espressione online. [16]

Una volta che il contenuto viene identificato come inappropriato, ci sono diverse possibili conseguenze, tra cui segnalazione o eliminazione del contenuto. In generale, le soluzioni pratiche spesso adottano un approccio "human-in-the-loop" che combina l'intervento umano con l'uso di algoritmi, in alcuni casi prevedendo la totale automazione per i casi più ovvi e l'intervento di moderatori umani per le situazioni complesse.

Di seguito un riassunto di vantaggi e svantaggi della moderazione automatizzata:

Vantaggi:

- È efficiente per grandi volumi di contenuti e può alleviare il carico di lavoro dei moderatori.
- Può essere applicata in tempo reale, risultando quindi più rapida.
- Può essere personalizzata per rilevare contenuti specifici.

Svantaggi:

- Gli algoritmi possono avere difficoltà a interpretare il contesto e le sfumature, portando a falsi positivi o falsi negativi.

- Gli algoritmi possono essere aggirati da contenuti sofisticati e ben studiati.

Ognuna di queste forme di moderazione presenta vantaggi e limiti e le piattaforme spesso combinano questi approcci per affrontare in modo efficace un'ampia gamma di contenuti. L'uso della moderazione automatizzata, ad esempio, può consentire una rapida identificazione di spam e contenuti ovviamente dannosi, ma può anche comportare il rischio di falsi positivi e di rimozione errata di contenuti legittimi. D'altro canto, la moderazione umana può fornire una valutazione più accurata di contenuti complessi, ma è soggetta a limitazioni in termini di scalabilità e tempi di risposta. In uno studio svolto nel 2023 sulla moderazione nella social VR [7] è stato osservato che su un totale di 100 eventi svoltisi in alcune delle piattaforme di social VR più popolari del momento (Rec Room, AltspaceVR e Horizon Worlds), nel 45% degli eventi sono stati praticati comportamenti dannosi da parte degli utenti e i moderatori erano presenti solo nel 51% degli eventi in cui sono stati osservati questi comportamenti. Inoltre, i moderatori non sono sempre intervenuti, probabilmente anche a causa di limitazioni fisiche e spaziali (vista ostacolata e audio spaziale).

Una delle sfide principali nella moderazione dei contenuti online è trovare un equilibrio tra la protezione degli utenti e la salvaguardia della libertà di espressione. Un'eccessiva censura o rimozione di contenuti può limitare la diversità di opinioni e il dibattito, minando così il principio fondamentale della libertà di espressione.

D’altro canto, una moderazione troppo morbida potrebbe consentire la diffusione di contenuti dannosi, abusi e incitamento all’odio, mettendo a rischio la sicurezza e il benessere degli utenti. Pertanto, trovare il giusto equilibrio è una sfida continua e richiede una continua adattabilità delle politiche e delle pratiche di moderazione.

Moderazione nella social VR

Molte delle sfide e delle pratiche di moderazione sviluppate nelle piattaforme più convenzionali sono applicabili anche nel contesto della social VR, ma alcune sfide uniche richiedono soluzioni specifiche. Infatti, queste esperienze coinvolgono gli utenti in ambienti virtuali tridimensionali, dove possono interagire, creare oggetti digitali e comunicare attraverso azioni in tempo reale.

La moderazione dei contenuti in queste esperienze è quindi una sfida complessa, che coinvolge non anche interazioni, comportamenti e comunicazione non verbale. La loro natura effimera e in tempo reale richiede un approccio di moderazione diverso dalle piattaforme convenzionali. Come visto poco fa, percezione e riposta dei moderatori possono essere limitate dalla spiccata fisicità degli ambienti virtuali in cui la vista può essere ostacolata e il suono si disperde con l’aumento della distanza dalla sorgente. Inoltre, le sfide tecniche nel distinguere tra contenuti dannosi e non dannosi possono causare problemi nella moderazione automatizzata.

Un altro fattore importante di cui tenere conto, è che le norme comportamentali variano a seconda del contesto di utilizzo; quindi, la moderazione dei contenuti nelle esperienze immersive richiede anche una comprensione approfondita del contesto e delle dinamiche sociali. Le piattaforme dovrebbero considerare il modo in cui le persone si comportano nei diversi contesti e adattare di conseguenza le politiche sui contenuti. Questo adattamento è complicato ovviamente dalla varietà delle situazioni. Le priorità e i meccanismi di applicazione delle politiche di moderazione possono variare notevolmente a seconda delle piattaforme e del contesto in cui vengono utilizzate. Ad esempio, in contesti familiari o individuali, le politiche enfatizzano la privacy e si basano su strumenti controllati dall'utente. Tuttavia, in situazioni in cui sono coinvolte molte persone, la moderazione può richiedere un approccio diverso, ad esempio la supervisione della comunità o dell'amministratore per bilanciare privacy e sicurezza. Quando si tratta di interazioni con estranei, la priorità è spesso la sicurezza e ci affidiamo alle segnalazioni degli utenti, agli strumenti automatizzati e al monitoraggio attivo per mantenere un ambiente sicuro. Mentre le comunicazioni private possono fare affidamento sulla moderazione basata sugli utenti, quelle pubbliche richiedono approcci più completi. È importante trovare un equilibrio tra sicurezza e privacy, consentendo agli utenti di personalizzare le proprie esperienze e mantenendo un ambiente aperto all'espressione creativa evitando comportamenti dannosi. [1]

Moderazione delle chat vocali

Nel caso specifico della moderazione delle chat vocali, questa presenta sfide uniche rispetto alla moderazione del testo.

Prima di tutto la gestione delle violazioni in questo contesto è complicata dalla natura in tempo reale delle chat vocali e dalla necessità di prendere decisioni immediate per mantenere l'integrità delle conversazioni. Come anticipato, un aspetto critico, che ritroviamo anche nelle chat testuali, è la soggettività e il contesto. Poiché i moderatori devono ascoltare le conversazioni vocali per valutare le violazioni delle regole, devono tenere conto del contesto in cui vengono pronunciati determinati termini o frasi. Ciò aggiunge complessità alle decisioni di moderazione, poiché la stessa parola potrebbe essere utilizzata in modi diversi a seconda del contesto.

La raccolta di prove è un'altra sfida fondamentale. A differenza dei messaggi di testo, le conversazioni vocali non sono facilmente documentabili o archiviabili e l'assenza di una traccia persistente delle interazioni rende difficile verificare e documentare le violazioni delle regole. Pertanto, i moderatori dovrebbero essere in grado di identificare le violazioni delle regole nel momento in cui si verificano o avere testimonianze e segnalazioni affidabili su cui basarsi per agire, ma questo non sempre è possibile. La natura effimera della social VR potrebbe rendere più

difficile anche il controllo di eventuale abuso di potere da parte dei moderatori stessi.

Un'ulteriore complicazione è il rumore, in particolare quando gli utenti producono intenzionalmente suoni forti o fastidiosi o contenuti audio inappropriati per disturbare gli altri e interrompere le conversazioni; a differenza delle chat testuali, dove il rumore è limitato al testo scritto, nelle chat vocali diventano rilevanti il volume e la qualità del suono. Il rumore può interrompere una conversazione in corso, rendendo difficile la partecipazione attiva, e rovinare l'esperienza degli utenti. La realtà virtuale può introdurre ulteriori sfide in questo contesto, in particolare nell'identificazione dei colpevoli. Infatti, a causa dell'utilizzo dell'audio spaziale, la voce di un utente può essere ascoltata solo da chi si trova nelle immediate vicinanze. Quindi, mentre da un lato è molto più difficile per un singolo utente interrompere l'esperienza di tutti gli altri, poiché non può essere sentito da tutti, a meno che non gli venga concesso l'accesso a qualche sorta di amplificazione vocale, allo stesso tempo ciò può ostacolare la moderazione.

Tutto ciò è complicato ulteriormente dalla diffusa mancanza di strumenti avanzati di moderazione audio automatica. Sebbene esistano e siano largamente utilizzati algoritmi di moderazione del testo in grado di rilevare parole chiave o frasi sospette, lo stesso non vale per gli strumenti equivalenti per l'audio. I moderatori devono

quindi monitorare costantemente le conversazioni vocali senza l'uso di automatismi.

Infine, la moderazione vocale richiede spesso decisioni personalizzate da parte dei moderatori. Poiché le regole possono consentire una certa flessibilità nell'interpretazione delle violazioni, i moderatori devono prendere decisioni in base al loro giudizio personale. Ciò rende la moderazione della chat vocale più soggettiva rispetto alla moderazione del testo, dove le violazioni possono essere valutate in modo più oggettivo sulla base delle parole scritte.

Piattaforme

La maggior parte delle piattaforme di social VR e dei videogiochi multiplayer online offre agli utenti sia degli strumenti che permettano loro di proteggersi autonomamente da eventuali molestie, sia strumenti di segnalazione che consentono loro di segnalare contenuti problematici, che verranno poi esaminati da moderatori umani, sia la possibilità di fare riferimento a moderatori umani. In alcuni casi vengono anche implementati sistemi automatici che monitorano i contenuti e le conversazioni degli utenti, ma nei contesti che abbiamo preso in considerazione questi agiscono prevalentemente su messaggi testuali, sia in maniera preventiva che reattiva, e solo in pochissimi casi la moderazione automatica viene applicata anche alle chat vocali con l'unico scopo di contrassegnare e proporre ai moderatori umani interazioni potenzialmente inappropriate in maniera automatica e quindi non preventiva.

In questa sezione vengono analizzate le piattaforme di social VR più popolari attualmente disponibili.

Rec Room

Rec Room [17] è una piattaforma social VR in cui gli utenti possono costruire, giocare e socializzare con amici da tutto il mondo, esplorare stanze create da altri giocatori, creare il proprio contenuto e personalizzare il proprio avatar. La maggior

parte dei contenuti di Rec Room sono generati dagli utenti, inclusi spazi, attività, oggetti e abbigliamento per avatar. La piattaforma ha un codice di condotta di base che i giocatori devono seguire e per mantenere la sicurezza nelle comunità si affida principalmente a moderatori volontari e alle segnalazioni degli utenti.

Codice di condotta

Il codice di condotta di Rec Room richiede ai giocatori di seguire determinate regole per garantire un ambiente positivo e rispettoso, vieta infatti l'utilizzo di linguaggio, comportamenti o contenuti sessisti, razzisti, discriminatori o molesti e la promozione di comportamenti illegali. Nelle stanze pubbliche non sono ammessi comportamenti che arrechino disturbo agli altri utenti, contenuti sessualmente espliciti o argomenti controversi. Questa regola non si applica solo alle stanze private o non elencate, ma è necessario che i proprietari di queste stanze si assicurino che tutti i presenti siano d'accordo con le attività svolte comunicando chiaramente il loro scopo agli utenti prima che essi entrino nella stanza. A differenza delle altre piattaforme, Rec Room non ha un limite minimo di età e, allo scopo di proteggere gli utenti più piccoli, i giocatori di età inferiore ai 13 anni devono utilizzare un account Junior.

Moderazione

Oltre a mettere a disposizione moderatori umani a cui fare riferimento all'interno delle comunità, Rec Room implementa una moltitudine di sistemi che consentono

ai giocatori di gestire situazioni specifiche in base alle loro preferenze personali e a ridurre al minimo gli incontri negativi:

Account Junior

Si tratta di account appositamente progettati per i giocatori di età inferiore ai 13 anni; questi account presentano diverse restrizioni, come la disabilitazione della chat vocale e testuale, l'assegnazione di uno username univoco privo di informazioni personali, la disabilitazione degli strumenti di creazione, tranne nella propria stanza privata, l'eliminazione del fuoco amico nelle missioni e l'accesso limitato ad eventi e club. Nel caso vengano rilevate attività tipicamente associate a minori dai sistemi di moderazione e verifica su un account non-junior, lo staff di Rec Room si riserva il diritto di convertirlo in un account junior. Inoltre, questa funzionalità permette ai giocatori non-junior di impostare il matchmaking in modo da evitare stanze o attività in cui sono presenti giocatori junior.

Matchmaking basato sull'età

Come accennato poco prima, i giocatori appartenenti alla stessa fascia di età vengono abbinati tra loro durante le partite, ciò significa che i giocatori saranno abbinati a persone che potrebbero avere interessi, esperienze e livelli di maturità simili, migliorando la qualità delle interazioni durante il gioco; inoltre, può contribuire a creare un ambiente di gioco più sicuro e appropriato per i giovani

giocatori, riducendo le possibilità di incappare in contenuti o comportamenti inadeguati.

Bolla personale

Ogni giocatore ha una bolla spaziale personale che può configurare liberamente, di default questa bolla è impostata a livello medio, ma è anche possibile aumentarne o diminuirne le dimensioni. Quando gli altri giocatori entrano nella bolla, diventano completamente invisibili. Questo sistema permette ai giocatori di avere un certo controllo sull'interazione con gli altri utenti definendo il raggio del loro spazio personale e riduce la possibilità di sentirsi a disagio o molestati da altri giocatori.

Mutare gli altri giocatori

Gli utenti possono silenziare singoli giocatori per evitare di sentirne l'audio durante il gioco se stanno disturbando o comportandosi in modo inappropriato. Possono anche silenziare l'audio della voce di tutti i giocatori o attivare l'audio della voce solo per gli amici.

Bloccare gli altri giocatori

I giocatori bloccati vengono automaticamente silenziati e svaniscono gradualmente e il sistema di matchmaking eviterà di abbinarli nuovamente all'utente bloccante così da evitare ulteriori interazioni negative o sgradevoli;

tuttavia, tramite invito di un amico, è ancora possibile entrare in una stanza pubblica in cui si trova il giocatore bloccato nel caso si desiderasse di partecipare a un'attività specifica.

Votazione

È possibile avviare una votazione per espellere un giocatore da un certo ambiente, tutti gli altri giocatori nella stanza verranno informati dell'avvio della votazione e verrà chiesto loro di votare sì o no. Se la maggioranza dei giocatori vota sì, il giocatore verrà rimosso dalla stanza e non potrà rientrare nella stessa sessione. Le votazioni per l'espulsione non richiedono una violazione del codice di condotta come avviene per le segnalazioni.

Segnalazioni

Gli utenti possono avviare una segnalazione se notano che un giocatore o una stanza sta violando il codice di condotta, verrà loro chiesto di selezionare la sezione del codice di condotta violata e fornire una breve descrizione testuale della violazione. In questo modo lo staff di Rec Room potrà successivamente prendersi carico della segnalazione e scegliere se e quali provvedimenti intraprendere.

Stanze pubbliche e private

Gli utenti possono creare un'istanza privata di qualsiasi attività di Rec Room o stanza. In una stanza privata, solo i giocatori invitati possono entrare. La coesistenza

di spazi pubblici e privati offre ai giocatori privacy, controllo e flessibilità nelle loro esperienze di gioco, permette loro di iniziare nel proprio spazio privato, invitare le persone che desiderano nei propri spazi e, allo stesso tempo, partecipare a stanze e attività pubbliche.

Impostazioni del microfono

Le impostazioni audio consentono ai giocatori di modificare il tono della propria voce o di passare alla modalità "push-to-talk" o di disattivare completamente il microfono, se lo desiderano. Ciò consente ai giocatori di decidere quando e come condividere la propria voce con gli altri e può essere particolarmente utile per coloro che desiderano mantenere la privacy.

Filtraggio di messaggi testuali e nomi

Rec Room utilizza tecniche di filtraggio per cercare di vietare e rimuovere eventuali frasi offensive o volgari presenti nei testi inviati dai giocatori, inclusi i nomi dei giocatori e i nomi delle stanze private.

Moderazione automatica delle chat vocali

Il team di Rec Room collabora con un sistema per rilevare automaticamente il linguaggio sessista, razzista, discriminatorio o violento all'interno dei giochi online. Per garantire la privacy dei giocatori, il sistema di moderazione vocale utilizza tecnologie avanzate che riducono al minimo, rendono anonimi ed eliminano i dati

vocali dei giocatori. Inoltre, il sistema è attivo esclusivamente nelle stanze pubbliche, consentendo ai giocatori di utilizzare stanze private per conversazioni completamente riservate.

La tecnologia alla base della moderazione automatica della chat vocale in Rec Room è alimentata da ToxMod, una soluzione innovativa di Modulate.AI. che utilizza l'intelligenza artificiale per rilevare e analizzare conversazioni tossiche in tempo reale. Attraverso modelli avanzati di apprendimento automatico, ToxMod comprende il contesto delle conversazioni, lo slang, le norme culturali. Questo approccio aiuta a prevenire comportamenti tossici nelle conversazioni vocali online riducendo la necessità di segnalazione dei giocatori o di monitoraggio casuale. Il funzionamento di ToxMod può essere suddiviso in tre fasi principali:

- Triage dei dati vocali: ToxMod valuta i dati delle conversazioni vocali, identificando quelli che richiedono indagini e analisi più approfondite. Questa fase di triage garantisce efficienza e precisione, eliminando silenzi o rumori di fondo irrilevanti.
- Analisi avanzata della tossicità: ToxMod esamina il tono, il contesto e l'intento nelle conversazioni filtrate, utilizzando processi avanzati di apprendimento automatico. Questa analisi valuta diversi elementi, come il tono della voce, l'emozione e il contesto, per determinare il tipo e la gravità del comportamento tossico.

- Gestione delle conversazioni tossiche: ToxMod identifica le conversazioni vocali più tossiche, consentendo ai moderatori di agire tempestivamente per mitigare comportamenti indesiderati. La console web di ToxMod fornisce informazioni dettagliate per ogni caso di comportamento tossico, consentendo ai moderatori di lavorare in modo efficiente anche con grandi volumi di conversazioni simultanee.

VRChat

Nel contesto della realtà virtuale sociale, VRChat [18] offre agli utenti la possibilità di plasmare la propria esperienza attraverso la creazione e la personalizzazione di avatar e mondi unici, utilizzando l'SDK di Unity o sfruttando le molteplici opzioni fornite dalla piattaforma stessa. La comunicazione in VRChat è basata sull'utilizzo di audio, chat, disegno e scultura, offrendo agli utenti diverse modalità di interazione. L'integrazione di giochi, spesso creati dalla community, come Capture the Flag e Battle Discs, arricchisce ulteriormente l'esperienza. L'esplorazione rappresenta un elemento cardine di VRChat, permettendo agli utenti di immergersi in centinaia di mondi creati da altri partecipanti, partecipare a eventi ufficiali e della comunità, e sviluppare nuove connessioni; il concetto di "istanze" si riferisce agli ambienti virtuali in cui gli utenti possono interagire e socializzare. Questi spazi possono essere pubblici o privati, con diverse modalità e regolamenti di accesso. Le istanze pubbliche sono aperte a tutti gli utenti e facilmente accessibili, mentre le

istanze private richiedono un invito o l'approvazione dell'organizzatore per accedervi. Le istanze di gruppo sono ambienti gestiti da specifici gruppi di utenti, ciascuno con le proprie regole e regolamenti interni.

Codice di condotta

Le linee guida comportamentali di VRChat incoraggiano a trattare gli altri con curiosità, rispetto e comprensione. L'utilizzo di VRChat o dei servizi forniti richiede un'età minima di 13 anni, con l'autorizzazione dei genitori richiesta per gli utenti di età compresa tra 13 e 17 anni. È vietata qualsiasi forma di linguaggio sessista, razzista, incitante all'odio, molesto o discriminatorio ed evidenziano la necessità di gestire i disaccordi in modo costruttivo, evitando commenti negativi ripetuti o gravi, nonché minacce o condivisione di informazioni personali. È vietata la promozione di comportamenti illegali, inclusa la violenza estrema. Le stesse regole devono essere applicate durante la creazione di contenuti all'interno della piattaforma. Le istanze pubbliche non devono essere disturbate da comportamenti o contenuti provocatori quali contenuti per adulti, discussioni controverse, attività sensibili o interruzioni dell'esperienza altrui, le istanze private devono seguire le regole stabilite dall'autorità presente, mentre le istanze di gruppo sono gestite autonomamente dai gruppi, con regole ragionevoli e conformi alle linee guida della comunità.

Moderazione

VRChat fornisce agli utenti diversi strumenti per moderare la loro esperienza nella piattaforma:

- **Mutare gli utenti**: è possibile silenziare gli altri utenti.
- **Bloccare gli utenti**: è possibile bloccare gli altri utenti, questi non saranno più in grado di vederti e tu non sarai più in grado di vederli e sentirli. Gli utenti bloccati possono comunque interagire con il mondo in modi che potrebbero essere percepiti.
- **Spazio personale**: si tratta di un'area che gli utenti possono attivare attorno a sé per far scomparire gli utenti nelle vicinanze, anche se possono comunque sentirli.
- **Segnalazione**: è possibile segnalare rapidamente gli altri utenti attraverso dei menù appositi o contattare il team di VRChat per segnalazioni più dettagliate.
- **Sistema di fiducia e sicurezza**: è possibile nascondere le caratteristiche degli avatar di altri utenti come suoni, immagini, animazioni e altro.

Sistema di fiducia e sicurezza

Il VRChat Trust and Security System è un peculiare strumento che espande il sistema di trust preesistente della piattaforma, con l'obiettivo di proteggere gli utenti da possibili fastidi causati da shader, suoni, effetti particellari e altri elementi

che potrebbero disturbare l'esperienza VRChat. Attraverso una valutazione di diversi parametri, come il tempo trascorso sulla piattaforma, le connessioni social stabilite e i contenuti generati dagli utenti, a ciascun partecipante viene assegnato un “livello di fiducia”, visibile sul nome utente quando si accede al menu rapido, che varia da “Visitatore” al "Leggendario". Il sistema di sicurezza consente di configurare le caratteristiche degli avatar di altri utenti in base al loro livello di fiducia. Attraverso diversi "livelli di scudo" che vanno da "Nessuno" a "Personalizzato", questo sistema può nascondere o mostrare avatar, icone, shader, particelle e luci degli utenti. La personalizzazione è ulteriormente migliorata dalla possibilità di escludere il sistema di sicurezza per singoli utenti o amici, consentendo un maggiore controllo su misura.

Meta Horizon Worlds

Meta Horizon Worlds [19] è una piattaforma di realtà virtuale che consente di esplorare e creare mondi virtuali e giochi multiplayer e di interagire con altri utenti in modalità cooperativa o competitiva.

Codice di condotta

Il codice di condotta di Meta sottolinea l'impegno nella creazione di comunità attraverso esperienze virtuali, promuovendo la libera espressione delle opinioni. Gli utenti sono tenuti a seguire i valori di Meta, quali sicurezza, autenticità, dignità e privacy. La distinzione tra esperienze chiuse e pubbliche attribuisce ai creatori la

responsabilità principale per la gestione di comportamenti e contenuti nelle esperienze chiuse, come i mondi riservati ai membri, ovvero spazi chiusi che consentono ai creator di creare e gestire un luogo in cui una community di persone può riunirsi. Sviluppatori, creatori e amministratori possono stabilire regole aggiuntive, definendo la cultura del proprio mondo e il contesto. La moderazione quotidiana dei mondi riservati ai membri è affidata al creatore e ai membri con ruolo di moderatore, con Meta garante dell'efficace moderazione da parte dei creatori e dell'intervento in situazioni di mancata conformità. Il codice vieta comportamenti illegali, abusi sessuali su minori, bullismo e violenza. Gli spazi pubblici sono soggetti a regole specifiche, vietando spam, contenuti sessuali o violenti e richiedendo il rispetto delle leggi locali e delle regole degli sviluppatori. Anche in questo caso l'età minima degli utenti è di 13 anni.

Moderazione

Nei mondi riservati ai soli membri, i creatori hanno la responsabilità primaria della moderazione e possono nominare moderatori e stabilire regole di moderazione. I report consentono ai creatori di monitorare il comportamento dei membri e intraprendere azioni conformi al codice di condotta. I moderatori hanno a disposizione alcuni strumenti di moderazione esclusivi che possono gestire per garantire un ambiente positivo e sicuro nei mondi riservati ai soli membri, come la possibilità di silenziare un utente per due ore in un mondo, impedendo a chiunque

di sentirlo o di rimuovere temporaneamente o permanentemente un utente dal mondo.

Altri strumenti di moderazione sono anche a disposizione di tutti gli utenti presenti in un mondo:

- **Area sicura**: si tratta di uno spazio personale in cui gli utenti possono staccarsi un attimo dalle altre persone a dagli ambienti circostanti e dal quale possono silenziare, bloccare e segnalare contenuti o persone.
- **Modalità voce**: questa funzione permette di non ascoltare le conversazioni di utenti estranei trasformando le loro voci in suoni incomprensibili ma comunque piacevoli.
- **Mutare gli utenti**: è possibile scegliere di mutare gli altri utenti, inoltre è disponibile una funzione di controllo livello utente che consente di silenziarli automaticamente.
- **Bloccare gli utenti**: è possibile bloccare gli altri utenti così da non sentirli e vederli più.
- **Segnalazione di mondi e utenti**: le segnalazioni di violazioni del codice di condotta vengono esaminate da un team di specialisti della sicurezza. Al team vengono forniti gli ultimi minuti di audio e altre interazioni su Horizon Worlds per valutare la situazione. Il contenuto segnalato viene eliminato una volta completata la verifica. In caso di segnalazioni provenienti da

esperienze pubbliche, gli specialisti possono monitorare la zona da remoto, intervenendo, ad esempio, rimuovendo l'utente o bloccandolo su Horizon Worlds.

- **Spazio personale**: questa funzione crea una barriera invisibile attorno all'avatar di un utente, impedendo ad altri avatar di avvicinarsi a meno di circa 1 metro di distanza. Gli utenti possono personalizzare questa funzione rendendola attiva solo per le persone che non seguono invece che per tutti.
- **Sondaggio**: questa funzione permette di chiedere in forma anonima ai membri di un gruppo se ritengono sia necessario allontanare una persona che disturba l'ambiente con comportamenti indesiderati. Dopo aver selezionato il motivo della rimozione, i membri possono votare a favore, contro o astenersi. Se la maggioranza è d'accordo sull'allontanamento, la persona viene espulsa dal mondo e trasportata nel suo spazio personale.
- **Account per minorenni**: gli account minorenni sono soggetti a restrizioni per garantire sicurezza e privacy, queste includono profili privati che richiedono l'approvazione di tutte le richieste dei follower, la possibilità di rendere visibili lo stato e la posizione solo su scelta degli utenti, la limitazione dell'interazione con adulti sconosciuti e l'esclusione dai mondi virtuali destinati agli adulti.

Mute assist

Successivamente all'esecuzione della ricerca e all'elaborazione di questa tesi e solo in determinati mondi in Meta Horizon Worlds, sono stati attivati sistemi

automatizzati finalizzati all'analisi dell'audio degli utenti per individuare espressioni volgari o parole potenzialmente offensive, con l'obiettivo di migliorare l'esperienza di coloro che preferiscono evitarle.

Il "Mute Assist" rappresenta una funzione di controllo dell'utente che consente di silenziare automaticamente o rapidamente tramite la comparsa di un pop up le persone sconosciute che utilizzano linguaggio volgare o parole potenzialmente offensive. All'ingresso in un mondo con "Mute Assist" attivo, gli utenti acconsentono alla revisione del proprio audio da parte di sistemi automatizzati, con la cancellazione dell'audio elaborato dopo il rilevamento. È possibile configurare la funzione scegliendo il livello di comfort preferito tra “spento”, “basso”, ovvero i termini volgari sono accettabili, ma l’utente ha l'opzione di silenziare le persone che utilizzano spesso parole potenzialmente offensive, “medio”, ovvero l’utente ha l'opzione di silenziare le persone che utilizzano ripetutamente parole volgari o potenzialmente offensive, e “alto”, ovvero vengono silenziate le persone che utilizzano termini volgari o offensivi anche solo in alcune occasioni.

Questionario preliminare

Per acquisire una comprensione approfondita delle percezioni e delle preferenze delle persone riguardo ai sistemi di moderazione online, è stato condotto un sondaggio completo. L'obiettivo dell'indagine era quello di ottenere una visione

completa di come gli utenti percepiscono tali sistemi e identificare gli aspetti che ritengono più rilevanti. La ricerca si è concentrata principalmente sulle comunità attive su piattaforme come Reddit e Discord, nello specifico su server e subreddit legati alla realtà virtuale e ai videogiochi. Un totale di 58 partecipanti hanno generosamente condiviso le loro opinioni attraverso il questionario, fornendo così un prezioso contributo all'analisi delle prospettive e delle esigenze degli utenti in relazione ai sistemi di moderazione online.

Risultati

Il campione dei partecipanti all'indagine risulta composto prevalentemente da giovani maschi bianchi, con un livello di istruzione medio-alto e residenti in Nord America o in Europa. Questo profilo può riflettere le caratteristiche della popolazione interessata alla realtà sociale virtuale e agli ambienti di gioco multiplayer online, ma può anche limitare la rappresentatività e la generalità dei risultati.

L'indagine mostra che gli ambienti di realtà virtuale sociale non sono ampiamente utilizzati dai partecipanti, con la maggioranza che afferma di non averli mai provati o di usarli raramente. Tra le piattaforme specifiche, VRChat sembra essere la più popolare, mentre le altre sono poco conosciute o frequentate. Molto più popolari tra i partecipanti risultano essere gli ambienti di gioco multiplayer online, con una

varietà di frequenze di utilizzo che vanno da occasionale a frequente. I giochi menzionati dai partecipanti appartengono a diverse categorie e generi, mostrando un'ampia gamma di interessi e preferenze di gioco.

L'indagine evidenzia che la lettura e il rispetto del codice di condotta delle piattaforme sono aspetti trascurati o variabili tra i partecipanti. Molti partecipanti non leggono mai o raramente il codice di condotta, mentre altri sono insicuri o non sempre o spesso lo rispettano. Ciò potrebbe indicare che i partecipanti non sono sufficientemente informati o consapevoli delle regole e delle linee guida delle piattaforme, o che non le considerano importanti o rilevanti

Emerge chiaramente che la molestia online è un fenomeno diffuso, con la maggior parte dei partecipanti testimonianti episodi di insulti personali, discorsi di odio e altre forme di abuso. Le funzionalità di auto moderazione, come il blocco degli utenti e la segnalazione dei comportamenti inappropriati, sono ampiamente utilizzate dagli utenti per gestire le molestie. L'impatto delle molestie sull'esperienza degli utenti varia, ma la consapevolezza del problema è generalmente alta. Le risposte indicano che la percezione dei partecipanti di essere supportati dalle piattaforme in cui si verificano le molestie è variabile; infatti, gli utenti ritengono che le piattaforme non stiano facendo abbastanza per prevenire e affrontare le molestie, sottolineando la necessità di azioni più decisive da parte delle piattaforme stesse. Questi dati evidenziano l’urgenza di affrontare il problema delle molestie online. È

essenziale che le piattaforme attuino politiche e misure di sicurezza più efficaci, con una risposta tempestiva alle segnalazioni degli utenti.

Dall'analisi dei dati raccolti attraverso il questionario emergono profonde considerazioni sulla moderazione online. La richiesta predominante è quella di una moderazione tempestiva ed efficiente. La ricerca della precisione nella moderazione evidenzia la necessità di algoritmi avanzati e moderatori competenti in grado di distinguere in modo intelligente tra le varie sfumature del contenuto. La trasparenza nelle politiche di moderazione è altrettanto essenziale, ma emerge una divergenza riguardo l'importanza della considerazione del contesto durante la moderazione. L'interesse per la capacità di contestare le decisioni di moderazione riflette la richiesta di un processo più giusto e accessibile. Anche il mantenimento della privacy durante la moderazione emerge come una questione critica, evidenziando la necessità di trasparenza nella gestione dei dati degli utenti. Infine, emerge una maggiore fiducia nei moderatori umani, nonostante l'aumento della moderazione automatizzata, che suggerisce un continuo apprezzamento per il discernimento umano, soprattutto in situazioni complesse. Come dimostrano questi risultati, la moderazione online deve bilanciare molteplici esigenze. La sicurezza è fondamentale, ma gli utenti vogliono anche giustizia, trasparenza e coinvolgimento attivo.

Tecnologie

Perspective API

La scelta dell'API da utilizzare per il rilevamento della tossicità è stata il risultato di una ricerca approfondita, volta a identificare tra le opzioni disponibili quelle accessibili e gratuite, in grado di classificare testo in lingua italiana e popolari nel panorama attuale. Dopo un'analisi approfondita, quattro API si sono distinte e sono state prese in considerazione per un'ulteriore valutazione approfondita. Le API coinvolte in questo processo decisionale includono OpenAI [20], Perspective API [21], Rewire e Azure AI Content Moderator [22].

È cruciale sottolineare che esistono limiti comuni a tutti questi modelli di intelligenza artificiale nel contesto della rilevazione di tossicità nei testi. Infatti, dal momento che i modelli di machine learning possono commettere errori, spesso si preferisce affiancare questi sistemi alla moderazione umana. Inoltre, i bias che si trovano nei dataset che vengono utilizzati per allenare i modelli possono avere un impatto sulle loro predizioni, per esempio potrebbero essere assegnati punteggi più alti a commenti contenenti termini riferiti a gruppi più frequentemente bersagliati (ad esempio parole come "nero", "musulmano", "femminista", "donna" o "gay"), poiché i commenti su quei gruppi sono sovra rappresentati nei commenti tossici presenti nei dati di allenamento. Non meno importante è il fatto che il punteggio che queste

API assegnano ad un commento indica la confidenza o la probabilità che il testo sia tossico in qualche modo, non la gravità, non si tratta quindi di una misura oggettiva.

Per raggiungere una decisione finale, sono state condotte una serie di test utilizzando diversi dataset di dati etichettati e soglie di tossicità differenti; con soglia di tossicità si intende il valore critico di probabilità o confidenza al di sopra del quale un messaggio è considerato tossico. Per ogni singola API sono stati misurati e valutati:

- **Veri positivi (VP)**: numero di volte in cui il modello identifica correttamente un elemento come tossico.
- **Veri negativi (VN)**: numero di volte in cui il modello identifica correttamente un elemento come non tossico.
- **Falsi positivi (FP)**: numero di volte in cui il modello identifica erroneamente un elemento come tossico quando in realtà non lo è.
- **Falsi negativi (FN)**: numero di volte in cui il modello non riconosce correttamente un elemento come tossico quando lo è.
- **Precisione**: percentuale di elementi identificati come tossici che sono effettivamente tali. Una precisione elevata indica che la maggior parte degli elementi identificati come tossici lo sono veramente, riducendo il numero di falsi positivi.

$$Precisione = \frac{VP}{VP + FP}$$

Un modello con una precisione alta evita di censurare o penalizzare messaggi innocui.

- **Recupero**: percentuale di elementi tossici che sono stati correttamente identificati dal modello. Un recupero elevato suggerisce che il modello è in grado di individuare efficacemente la maggior parte degli elementi effettivamente tossici, riducendo il numero di falsi negativi.

$$Recupero = \frac{VP}{VP + FN}$$

Un modello con un recupero alto evita di ignorare o tollerare messaggi dannosi.

- **Tempo di risposta**: tempo che l'API impiega per elaborare e restituire una risposta.

Queste metriche hanno fornito un quadro completo delle prestazioni di ciascuna API in diverse situazioni. Dopo essere stati costretti in corso d'opera a rinunciare al potenziale utilizzo di Rewire a causa della sua dismissione, alla luce di questi risultati e consapevoli dei limiti intrinseci dei modelli utilizzati, che rappresentano una sfida non completamente risolvibile, la scelta finale è ricaduta su Perspective API. La selezione dell'API più adatta è stata guidata dalla necessità di garantire un rilevamento sufficientemente accurato dei contenuti tossici, ma il più bilanciato possibile e con una risposta efficiente e tempi di elaborazione accettabili, e dalle

limitazioni imposte dai vari servizi, che nel caso di Perspective si limitano a poter inviare solo 1 QPS.

	VP	FP	VN	FN	Precisione	Recupero	Tempo di risposta
Perspective	135	50	689	125	0,76	0,52	121,05 ms
OpenAI	140	54	700	106	0,72	0,56	603 ms
Azure	217	255	462	41	0,46	0,84	152,23 ms

Tabella 1 Media delle misurazioni effettuate durante i test.

Perspective è un'API gratuita nata dalla collaborazione tra Jigsaw e il gruppo Counter Abuse Technology di Google. Rappresenta il risultato di un'iniziativa di ricerca denominata Conversation-AI, volta a "esplorare le potenzialità e i limiti dell'apprendimento automatico nel contrastare la tossicità e le molestie online" [23]. Questa API utilizza l'apprendimento automatico per identificare un insieme di "concetti emotivi", chiamati attributi, relativi a un testo (ad esempio se il testo è tossico, minaccioso, offensivo, ecc.) e valuta l'impatto che potrebbe avere su una conversazione assegnando a ciascun attributo un punteggio compreso tra 0 e 1. Questo punteggio indica la probabilità che il lettore percepisca il testo come appartenente alla categoria dell'attributo corrispondente. I modelli sono stati addestrati utilizzando milioni di commenti provenienti da varie fonti online, tra cui Wikipedia e il New York Times, in più lingue. Questi commenti sono stati etichettati da 3-10 valutatori umani e il punteggio per ciascun attributo è stato calcolato come il rapporto tra i valutatori che hanno contrassegnato il commento. Gli sviluppatori

indicano che uno dei principali obiettivi futuri è integrare il contesto conversazionale nel processo di classificazione.

Gli attributi disponibili sono:

- **TOXICITY**: commenti offensivi, mancanti di rispetto o irragionevoli che possono spingere le persone a lasciare una discussione.
- **SEVERE_TOXICIT**Y: commenti estremamente odiosi, aggressivi o mancanti di rispetto, molto probabile che inducano un utente a lasciare una discussione. Meno sensibile a forme più lievi di tossicità.
- **IDENTITY_ATTACK**: commenti negativi o odiosi rivolti a qualcuno a causa della sua identità.
- **INSULT**: commenti offensivi o negativi rivolti a una persona o a un gruppo.
- **PROFANITY**: uso di parolacce, espressioni volgari o altro linguaggio osceno o profano.
- **THREAT**: commenti che esprimono l'intenzione di infliggere dolore, lesioni o violenza a un individuo o a un gruppo.

È possibile indicare esplicitamente nella richiesta gli attributi che si vogliono valutare e ricevere così un punteggio per ogni attributo.

Microsoft Azure Speech Service e SDK

Il servizio vocale di Microsoft Azure [24] è un servizio cloud avanzato che consente l'integrazione di potenti funzionalità di sintesi vocale e di riconoscimento nelle applicazioni. Questo servizio, basato su modelli linguistici avanzati e tecnologie di intelligenza artificiale, offre una notevole precisione nella trascrizione e nella sintesi vocale. L'SDK associato semplifica ulteriormente l'integrazione di queste funzionalità nei progetti, offrendo supporto per vari linguaggi di programmazione come C#, Python, Java e altri. Gli sviluppatori possono sfruttare questo SDK per implementare funzionalità avanzate, tra cui il riconoscimento vocale in tempo reale, la traduzione automatica della lingua e la personalizzazione del modello linguistico.

RiveScript

RiveScript [25] è un linguaggio di scripting progettato per la creazione di chatbot e interfacce di conversazione. È stato sviluppato per essere facile da imparare e da utilizzare, consentendo agli sviluppatori di creare rapidamente conversazioni complesse e interattive. Il linguaggio si basa su un formato di script semplice e leggibile, dove si possono definire regole di conversazione, risposte e comportamenti del chatbot. RiveScript supporta una vasta gamma di funzionalità, inclusi modelli di conversazione condizionali, variabili, possibilità di gestire il contesto delle conversazioni e altro ancora. Di seguito alcuni concetti base di RiveScript:

- **Regole di Conversazione**: in RiveScript, le conversazioni sono organizzate in regole. Una regola è composta da un "trigger" e dalle risposte associate. Il trigger è ciò che attiva una regola, e le risposte sono ciò che il chatbot dovrebbe rispondere a quel trigger. Ad esempio:

```
+ ciao come stai
- Bene, e tu?
- Tutto bene!
- Bene :) tu?
- Alla grande! Tu?
- Sto bene, grazie!
```

In questo caso il chatbot risponderà alla domanda “Ciao, come stai?” selezionando una risposta casuale tra quelle elencate nel trigger.

- **Wildcards**: i trigger possono includere wildcard per catturare varie parti dell’input. Ad esempio:

```
+ mi chiamo *
- Piacere di conoscerti, <star>!

+ preferisci * o *
- <star1> o <star2>, non fa nessuna differenza.
```

In questo caso il primo trigger catturerà frasi come "Mi chiamo Marco" o "Mi chiamo Sofia", consentendo al chatbot di memorizzare il nome dell'utente in delle variabili e utilizzare questa informazione nelle interazioni successive.

- **Variabili**: RiveScript supporta l'uso di variabili per memorizzare e recuperare informazioni durante una conversazione. Ad esempio:

```
+ mi chiamo *
- <set name=<star>> Piacere di conoscerti,<get name>!
```

In questo caso la wildcard viene catturata e memorizzata nella variabile ‘name’ così da poter accedervi in qualsiasi momento.

- **Arrays**: è possibile creare degli array di stringhe e accedervi dal trigger, questa funzione risulta molto utile per creare delle liste di sinonimi e ampliare lo spettro di azione dei trigger.

```
! array colori = rosso rossa verde blu giallo gialla

+ ti piace la mia maglietta (@colori)
- Certo, mi piace molto!
```

In questo caso il trigger catturerà frasi come “Ti piace la mia maglietta rossa?” o “Ti piace la mia maglietta blu?”

- **Risposte condizionali**: le risposte possono includere condizioni per rendere le interazioni più dinamiche. Ad esempio:

```
+ come mi chiamo
* <get name> == undefined => Non mi hai ancora detto il tuo nome
- Ti chiami <get name>, giusto?
```

In questo caso il chatbot darà due risposte diverse in base al valore della variabile ‘name’.

- **Topics**: RiveScript è in grado di mantenere uno stato di "contestualità" durante la conversazione. Il contesto di default è chiamato ‘global’.

```
+ ti odio
- Sei stato scortese, adesso scusati. {topic=scuse}

> topic scuse
      + ti odio
      - Va bene, ti perdono. {topic=global}

      + *
      - Non ti parlerò finché non ti sarai scusato!

< topic
```

Esistono interpreti RiveScript per diversi linguaggi di programmazione, nel nostro caso è stato integrato un interprete C# [26] direttamente nell’applicazione finale.

Unity

Unity [27] è un popolare game engine che ha conquistato una vasta adozione nell'industria del software, in particolare nello sviluppo di videogiochi e applicazioni in tempo reale. Grazie all'editor dedicato e all'esistenza di numerosi tool e risorse, Unity consente di creare esperienze interattive per diverse piattaforme, tra cui PC, console, dispositivi mobili, realtà virtuale e realtà aumentata. Unity ha una comunità ampia e attiva che condivide risorse, soluzioni e conoscenze, rendendo più accessibile l'apprendimento e l'uso della piattaforma.

VRoid Studio

VRoid Studio [28] è un software di creazione di personaggi 3D sviluppato da Pixiv per consentire agli utenti di progettare facilmente avatar personalizzati a partire da dei modelli 3D di base. VRoid Studio offre una vasta gamma di strumenti che consentono agli utenti, anche senza esperienza nella modellazione 3D, di creare personaggi unici e dettagliati. Gli utenti possono regolare aspetti come la forma del viso, i capelli, gli occhi, i vestiti e molti altri dettagli. Il software è diventato popolare tra artisti digitali, creatori di contenuti e sviluppatori che desiderano incorporare personaggi originali nei loro progetti in modo efficiente. VRoid Studio è ampiamente utilizzato anche nella comunità di VRChat per la creazione di avatar personalizzati.

Blender

Blender [29] è un potente software open-source per la modellazione, l'animazione, il rendering e la creazione di contenuti 3D. Grazie alla sua versatilità e alle numerose funzionalità integrate, Blender è ampiamente utilizzato sia da professionisti del settore che da appassionati di computer grafica. Blender consente di creare modelli complessi, animazioni dettagliate e scene realistiche, mette a disposizione un editor per creare materiali avanzati, un sistema di animazione completo, il supporto per il texturing e il rendering in tempo reale.

Prima di importare gli avatar creati con VRoid Studio in Unity, è spesso necessario ottimizzarli per garantire prestazioni fluide all'interno dell'ambiente virtuale. In questo contesto, gli add-on di Blender giocano un ruolo fondamentale poiché semplificano il processo di ottimizzazione e adattamento degli avatar. Questi includono funzionalità come la riduzione automatica della complessità della geometria, la correzione delle texture e l'organizzazione dei materiali, inoltre permettono l'integrazione di componenti necessari per l'interazione avanzata in piattaforme come VRChat, tra cui l'eye tracking e l'animazione delle espressioni facciali. I seguenti add-on sono stati utilizzati per lo sviluppo dell'applicazione:

VRM Blender Add-on

Questo add-on [30] è progettato per semplificare e ottimizzare il processo di creazione, modifica e preparazione di modelli 3D compatibili con il formato VRM (Virtual Reality Model), cioè il formato in cui vengono esportati i modelli da VRoid Studio.

Cats Blender Add-on

Questo plugin [31] offre una serie di funzionalità utili per gli utenti che lavorano con modelli 3D destinati a piattaforme come VRChat. Le sue funzionalità spaziano dalla correzione automatica della geometria e delle texture al supporto per la creazione di modelli di espressioni facciali avanzate e alla semplificazione del processo di conversione dei modelli in formato compatibile con VRM, molto usato nelle applicazioni di realtà virtuale.

Material Combiner Blender Add-on

Questo add-on [32] è uno strumento per combinare materiali e texture in Blender, gli utenti possono personalizzare dimensioni e colori, creare livelli per ogni immagine e impacchettare le UV in modo efficiente.

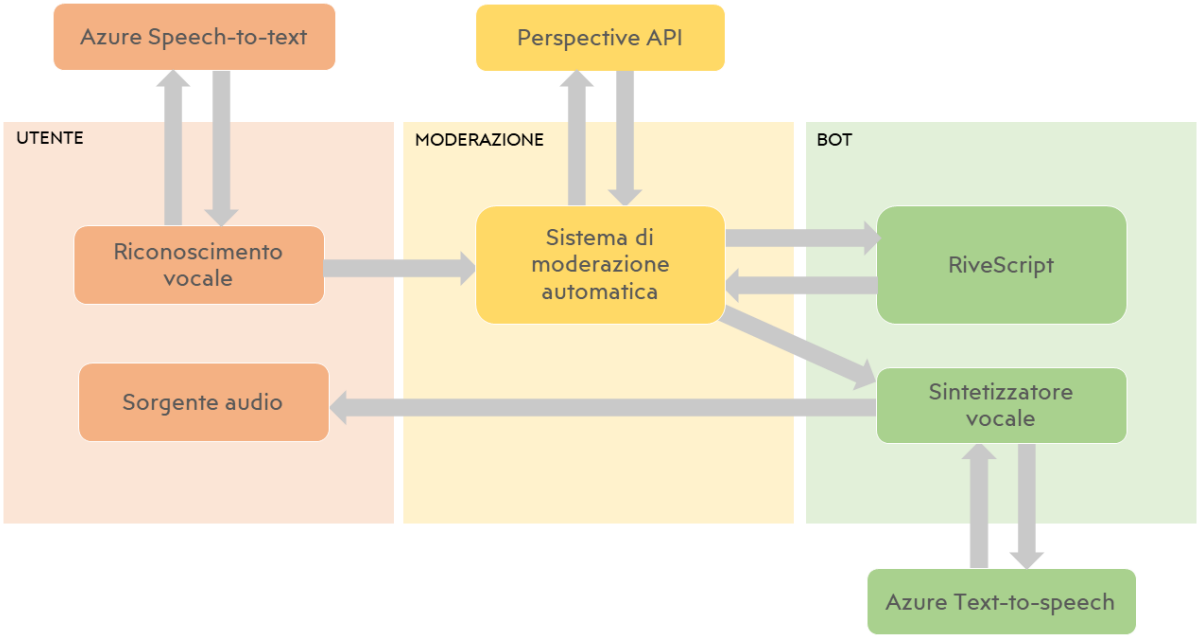
Mixamo

Mixamo [33] è una piattaforma online che permette di animare personaggi 3D in modo facile e veloce. La caratteristica distintiva di Mixamo risiede nella sua vasta libreria di animazioni predefinite pronte per l'uso, che coprono un'ampia gamma di movimenti e azioni umani. Gli utenti possono caricare un modello 3D e applicare facilmente le animazioni desiderate dalla libreria di Mixamo. Inoltre, Mixamo offre strumenti di rigging automatico, semplificando il processo di preparazione del modello per l'animazione. Questa piattaforma è particolarmente apprezzata nel contesto della creazione di contenuti per videogiochi, simulazioni virtuali e animazioni digitali, poiché offre una soluzione efficiente e accessibile per arricchire i progetti con movimenti realistici e dinamici.

Sviluppo software

L'applicazione concepita per le sperimentazioni della tesi offre uno scenario di simulazione di un ambiente sociale in realtà virtuale. Questa piattaforma mette a disposizione degli utenti un unico spazio tridimensionale di muoversi liberamente in prima persona e impegnarsi in conversazioni con gli altri presenti. Gli utenti sono dotati di strumenti di moderazione comunemente presenti delle piattaforme di social VR, ovvero la capacità di mutare, bloccare e segnalare altri utenti, oltre alla possibilità di attivare una bolla spaziale per proteggere il proprio spazio personale. Oltre agli strumenti di moderazione manuali, un sistema di moderazione automatica basato su Perspective API è stato integrato per identificare e bloccare in automatico i messaggi ritenuti inappropriati. Trattandosi di una simulazione, gli altri utenti sono interpretati da una serie di avatar virtuali che possono interagire con l'utente; nonostante si tratti di bot, le loro capacità non si limitano a poter ascoltare quello che l'utente dice e a rispondergli; infatti, i bot hanno accesso a loro volta agli strumenti di moderazione e possono approcciare il giocatore di loro spontanea volontà. Di seguito uno schema dell'architettura del software.

Figura 1 Schema generale dell'architettura del sistema di moderazione.



Interfaccia grafica

Menu principale

Attraverso il menu principale l'utente può accedere alle impostazioni, accedere alla lista degli utenti presenti nella stanza e uscire dall'applicazione.

Figura 2 Menu principale

Pannello delle impostazioni

Attraverso il pannello delle impostazioni è possibile attivare e disattivare la bolla personale e cambiarne la dimensione.

Figura 3 Pannello delle impostazioni

Pannello degli utenti presenti nella stanza

Attraverso questo pannello è possibile visualizzare una lista degli utenti presenti nella stanza e accedere rapidamente ad alcuni degli strumenti di moderazione a disposizione; infatti, accanto al nome di ogni utente sono presenti tre bottoni, uno per mutarlo, uno per bloccarlo e uno per segnalarlo, e un'icona che mostra lo stato dell'utente, ovvero se questo è bloccato o mutato.

Figura 4 Lista degli utenti presenti nella stanza

Pannello dei singoli utenti

Questo pannello, a cui si accede cliccando direttamente su un altro avatar, permette di accedere agli strumenti di moderazione che l'utente ha a disposizione e di intraprendere un'azione nei confronti dell'avatar in questione. Il pannello presenta tre bottoni, uno per mutare, uno per bloccare e un per segnalare l'avatar.

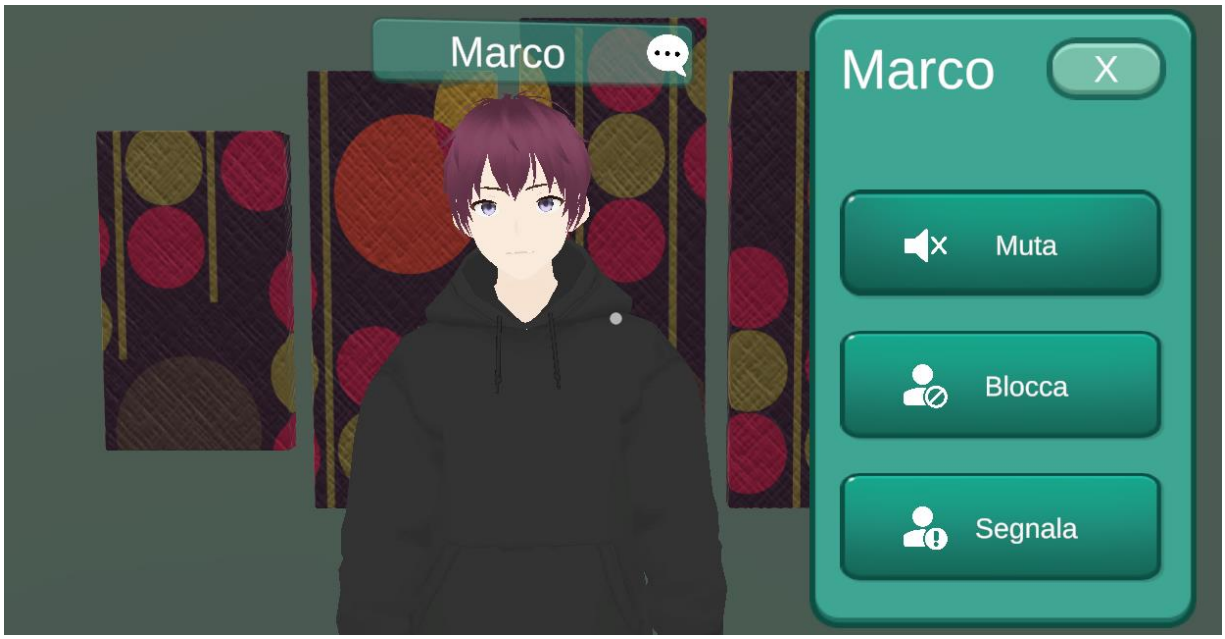


Figura 5 Pannello dei bot

Modulo per la segnalazione

I campi presenti in questo pannello devono essere compilati dall'utente per inviare una segnalazione; è necessario scegliere la categoria in cui rientra il motivo della segnalazione ed è possibile aggiungere commenti. Una volta compilato il modulo basta cliccare sul tasto di conferma per inviare la segnalazione. È fondamentale sottolineare che, sebbene l'utente riceva una notifica dell'avvenuta segnalazione, in realtà non viene effettivamente inviata alcuna segnalazione. Questa simulazione è stata progettata per riflettere un aspetto critico dell'esperienza di moderazione online. Su molte piattaforme, i moderatori sono spesso oberati di lavoro e il tempo trascorso prima che una segnalazione venga presa in considerazione può essere significativo. Di conseguenza, l'utente riscontra un ritardo nel ricevere una risposta, proprio come accade nella simulazione, dove l'utente non riceverà mai la risposta al report.

Figura 6 Modulo per la segnalazione

L’utente

Il sistema utente di questa applicazione permette di esplorare l'ambiente virtuale in prima persona e di interagire con i bot utilizzando il riconoscimento vocale, ovvero attraverso il semplice utilizzo della voce.

Input e controlli

Il movimento all'interno di questo spazio virtuale è stato implementato con l'utilizzo del "First Person Character Controller" incluso negli standard assets di Unity. Questo controller offre una solida base per il movimento in prima persona, garantendo una navigazione agevole e realistica. La scelta di utilizzare uno standard asset ampiamente adottato contribuisce a garantire una coerenza nella risposta ai comandi e una certa familiarità per gli utenti. L’utente ha anche la possibilità di aprire i menù accessibili durante la simulazione, sia cliccando sugli altri avatar che utilizzando delle shortcut apposite.

Riconoscimento vocale

Il riconoscimento vocale è stato implementato appoggiandosi all'SDK fornita da Microsoft per accedere allo Speech Service di Azure. All'interno della scena virtuale, è stato posizionato un oggetto dedicato che rimane costantemente in ascolto di un eventuale input dal microfono dell'utente. Una volta rilevato un input vocale,

viene attivato il processo di riconoscimento vocale fino a che l'utente non smette di parlare e la trascrizione testuale risultante viene poi inviata al sistema di moderazione automatica per la classificazione. In questo caso, la trascrizione ricevuta dal sistema di moderazione contiene tutto il contesto della frase pronunciata dall'utente, al contrario di quello che succederebbe se venissero inviate trascrizioni parziali man mano che l'utente parla; questo ridurrebbe la latenza del sistema, permettendogli di intervenire prima che l'utente smetta di parlare, ma potrebbe ridurre la capacità del sistema di comprendere il contesto della frase, aumentando la possibilità di falsi positivi. In futuro si potrebbe ricercare un livello di frammentazione dell'input dell'utente che offra un valido compromesso tra la riduzione della latenza e il mantenimento dell'accuratezza del sistema.

I bot

Nel contesto di questa applicazione, un elemento centrale è rappresentato dalla presenza di quattro chatbot distinti all'interno dell'ambiente virtuale. Ognuno di questi agenti virtuali è caratterizzato da un aspetto unico, un modo di comportarsi distintivo e un approccio specifico nelle interazioni con gli utenti. Questa diversità di personalità e stili di conversazione ha lo scopo di creare un'esperienza di simulazione sociale in realtà virtuale ancora più ricca e coinvolgente e che proponga una varietà di situazioni possibili. Inoltre, ogni bot è in grado di muoversi autonomamente all'interno dell'ambiente virtuale e di approcciare l'utente di loro iniziativa con una certa probabilità. Infine, anche i bot hanno accesso agli strumenti di moderazione e possono intraprendere delle azioni nei confronti dell'utente. Nei seguenti capitoli, esploreremo come questi chatbot siano integrati nell'ecosistema dell'applicazione.

Avatar

Gli avatar sono stati creati utilizzando VRoid Studio, che ha consentito di definire un aspetto unico per ciascun avatar. Per conferire loro vita, sono stati animati utilizzando la piattaforma online Mixamo. Quest'ultima, mettendo a disposizione una vasta libreria di animazioni predefinite, ha velocizzato il processo di animazione degli avatar.

Chatbot

Il sistema di chatbot all'interno di questa simulazione poggia su RiveScript, grazie al quale è stato possibile definire il comportamento peculiare di ciascun chatbot, conferendogli personalità uniche e approcci unici nelle interazioni con gli utenti. Prima di elaborare l'input dell'utente tramite Rivescript per selezionare una risposta, il messaggio passa attraverso un processo di preelaborazione. Durante questo passaggio, la punteggiatura viene rimossa, ogni carattere viene reso minuscolo e ogni parola viene sostituita con la parola più simile presente nello script RiveScript. Questo approccio mira a mitigare potenziali errori o variazioni linguistiche dovute alla grammatica o al genere che potrebbero impedire a RiveScript di identificare un trigger appropriato. Inoltre, questa procedura facilita lo sviluppo consentendo la creazione di elenchi di sinonimi più concisi senza la necessità di includere tutte le possibili coniugazioni verbali o variazioni di genere dei sostantivi. Per chiarire ulteriormente questo processo, è opportuno notare che per RiveScript, input come "Sono tuo amico" e "Sono tua amica" non sono considerati equivalenti, e per questo non vengono catturati dallo stesso trigger, a meno che non vengano definiti e utilizzati array che includono sia la forma maschile che quella femminile di ciascun termine. Per valutare la somiglianza delle stringhe durante questo processo, è stato utilizzato l'algoritmo di Jaro. L'algoritmo Jaro è una misura della somiglianza tra due stringhe che tiene conto delle caratteristiche comuni e della sequenza di caratteri. In sostanza, l'algoritmo assegna un punteggio in base al numero di caratteri

comuni e alla loro posizione relativa nelle due stringhe. Più alto è il punteggio Jaro, maggiore è la somiglianza tra le stringhe.

Sintetizzatore vocale

Ogni bot ha integrato al suo interno un sintetizzatore vocale. La tecnologia di sintesi vocale si basa su Azure Speech Service di Microsoft, che consente di trasformare le risposte testuali dei chatbot in audio fluido e naturale. Quando un chatbot formula una risposta o intraprende un'interazione spontanea, il testo viene inviato al servizio di sintesi vocale, che restituisce un file audio contenente la risposta del bot. Questo audio viene quindi riprodotto per consentire all'utente di ascoltare la risposta.

Caratterizzazione

Di seguito vengono illustrati la personalità e il comportamento di ogni bot

Figura 7 I bot presenti nella simulazione: Sofia, Marco, Luca e Alessia

Sofia

Sofia, il primo bot di questa simulazione virtuale, si presenta come una giovane diciottenne, irradiante di energia vivace e un sorriso contagioso. La sua personalità è intrisa di socievolezza e apertura, creando un'atmosfera accogliente per chiunque si trovi in sua compagnia. Caratterizzata da una straordinaria empatia, Sofia si dedica al benessere degli altri e ha sempre uno sguardo positivo sulle situazioni.

Le sue passioni abbracciano l'arte della musica in tutte le sue sfumature. Chitarrista di talento e appassionata di rock, la sua mente aperta la porta ad apprezzare una vasta gamma di generi ed esplorare nuovi artisti. Sofia è anche un'appassionata esploratrice culinaria, apprezza i piaceri gastronomici provenienti da ogni angolo del mondo e condivide la sua passione con gli altri. Nonostante la sua natura ottimista, Sofia mantiene una ferma determinazione. La sua tolleranza agli insulti ripetuti è limitata e, sebbene sia sincero, blocca l'utente di fronte agli insulti persistenti.

Marco

Marco, il secondo chatbot di questa esperienza virtuale, si presenta come un ragazzo di 17 anni dalla personalità estremamente prepotente e arrogante. La sua autostima si esprime attraverso atteggiamenti provocatori e maleducati, soprattutto sui social, dove trova soddisfazione nel lanciare duri insulti agli altri

utenti solo per suscitare reazioni negative. La sua propensione a infastidire gli altri evidenzia una mancanza di empatia e una scarsa considerazione per gli effetti delle sue azioni sulle persone che lo circondano. Un aspetto distintivo del carattere di Marco è la sua reazione alle interazioni con gli utenti. Dopo un numero prestabilito di interazioni, Marco comincia a seguire l'utente e ad insultarlo insistentemente. Questo comportamento aggiunge un elemento di sfida e difficoltà all'esperienza, sia in termini di molestie verbali che di invasione dello spazio personale dell'utente.

Nonostante la sua sfrontatezza, anche Marco, come Sofia, ha i suoi limiti. Se sottoposto lui stesso ad insulti, Marco blocca l'utente, imponendo una sorta di disciplina virtuale. In questo modo, l'applicazione non solo esplora le complesse dinamiche delle interazioni online, ma introduce anche un elemento di responsabilità e conseguenza all'interno del mondo virtuale.

Luca

Luca, il terzo bot di questa esperienza virtuale, si presenta come un giovane ragazzo di 15 anni, noto per il suo carattere eccezionalmente amichevole e scherzoso. Nel gruppo è considerato l'amico per eccellenza, sempre pronto a scherzare o a rompere il ghiaccio nelle situazioni sociali. La sua comunicazione è disinibita e spesso arricchita da espressioni un po' lascive o scherzose, con l'unico intento di divertire gli altri senza voler mai offendere o ferire i sentimenti di nessuno. Luca è un

appassionato di videogiochi, con una predilezione per i giochi battle royale e sparatutto.

La presenza di Luca all'interno dell'ambiente virtuale aggiunge una dimensione giocosa e amichevole alle interazioni degli utenti, creando un contesto in cui l'uso di espressioni scherzose, seppur un po' ardito, non ha alcun intento malevolo di fondo.

Alessia

Alessia, l'ultimo bot di questa simulazione virtuale, è una giovane ragazza di 20 anni, la cui personalità può essere descritta come invadente e fastidiosa. Utilizza i social media come terreno fertile per i suoi persistenti tentativi di flirtare con altri utenti. La sua natura civettuola e le sue tendenze a fischiare possono farla sembrare inappropriata nelle conversazioni, spesso ignorando i confini personali degli altri. L'interesse predominante di Alessia sembra concentrarsi sulla sfera sessuale e sulla sperimentazione delle dinamiche sociali online. La sua continua ricerca di approvazione attraverso comportamenti provocatori può spingerla ad andare oltre i limiti altrui, generando interazioni che possono risultare problematiche.

La presenza di Alessia all'interno della simulazione virtuale aiuta ad esplorare le complesse dinamiche delle interazioni online, evidenziando le sfide legate all'approccio invasivo e provocatorio nelle interazioni sociali digitali.

Il sistema di moderazione

Come anticipato, la moderazione è basata su una serie di azioni che gli utenti possono intraprendere nei confronti degli altri avatar e un sistema di moderazione automatica che blocca preventivamente i messaggi classificati come inappropriati. Analizziamo adesso nel dettaglio le diverse parti.

Mutare

L'utente ha la possibilità di mutare un singolo utente in modo da non sentire più quello che dice. L'utente mutato può ancora sentire quello che l'altro utente dice. È possibile farlo cliccando sul bot che si vuole mutare o attraverso la lista degli utenti presenti nella stanza. Il processo per riattivare l'audio di un utente è analogo.

Figura 8 Pop up che segnala che l'azione intrapresa dall'utente ha avuto successo.

Bloccare

L'utente ha la possibilità di bloccare un altro utente in modo da non vederlo più e non sentire più quello che dice, senza che per questo venga rimosso dalla stanza. Anche l'utente bloccato non potrà più vedere né sentire l'altro utente. È possibile farlo cliccando sul bot che si vuole mutare o attraverso la lista degli utenti presenti nella stanza. Per sbloccare un utente bloccato è necessario accedere alla lista degli utenti presenti nella stanza.

Segnalazioni

È possibile segnalare un utente cliccando sul suo avatar o accedendo alla lista degli utenti presenti nella stanza. Una volta cliccato sul bottone apposito, si apre un modulo che è necessario compilare indicando la motivazione della segnalazione. Come già spiegato, questo processo non ha conseguenze reali, poiché nessuna segnalazione verrà inviata, ma, poiché l'utente non riceverà mai un resoconto della propria segnalazione, riflette le tempistiche dilatate che spesso sono richieste su queste piattaforme prima che i moderatori si prendano carico della segnalazione e intraprendano delle azioni nei confronti dell'utente segnalato.

Bolla personale

È possibile attivare una bolla personale, ovvero un'area attorno al proprio avatar all'interno della quale gli altri utenti diventano invisibili e vengono mutati, senza che vengano bloccati permanentemente. Questo permette all'utente di proteggere il proprio spazio personale nel caso ne senta il bisogno. Attraverso il pannello delle impostazioni è possibile attivare o disattivare la bolla e cambiarne le dimensioni avendo a disposizione tre opzioni, ovvero "piccola", "media" e "grande".

Sistema di moderazione automatica

Il sistema di moderazione automatica si interfaccia direttamente con l'utente attraverso il riconoscimento vocale e con i bot.

Non appena il sistema di riconoscimento vocale completa la trascrizione dell'audio dell'utente, la invia al sistema di moderazione; questo mette in coda una richiesta contente la trascrizione appena ricevuta che verrà poi inviata in maniera asincrona a Perspective API man mano che il sistema evade i task di moderazione automatica. Non appena il sistema riceve in risposta alla richiesta una classificazione da parte dell'API, controlla lo score di ogni attributo e lo confronta con la soglia di tossicità: se lo score supera la soglia, blocca il messaggio e mostra un pop up per avvisare l'utente che ha infranto le regole e che per questo motivo il suo messaggio è stato bloccato automaticamente, altrimenti invia il messaggio al bot target

dell'interazione. Analogamente, quando un bot genera una risposta o effettua un'interazione spontanea, il messaggio viene prima inviato al sistema di moderazione per la classificazione: anche in questo caso, se lo score di uno degli attributi supera la soglia di tossicità, il messaggio viene bloccato e viene mostrato un pop up all'utente per avvisarlo che il bot ha infranto le regole e per questo motivo il suo messaggio è stato bloccato automaticamente, altrimenti il messaggio viene inviato al sintetizzatore del rispettivo bot al fine trasformare il messaggio nell'audio che verrà poi fatto sentire all'utente.

Sperimentazioni

Come già anticipato, lo scopo di questo studio è condurre una valutazione comparativa tra i metodi tradizionali di moderazione e un sistema di moderazione automatica in un ambiente di social esaminando come questi diversi approcci influiscano sulla percezione e sull'esperienza dei partecipanti in termini di interazioni sociali, sicurezza, libertà di espressione e naturalezza delle conversazioni.

Partecipanti

I partecipanti a questo studio sono stati selezionati in modo da garantire una ricerca etica e sicura, per questo sono stati reclutati solo adulti di almeno 18 anni di età e, e in modo da rendere il campione variegato per quanto possibile. In totale hanno eseguito il test 13 persone di età variabile tra i 18 e i 65 anni, di cui quasi la metà concentrata nella fascia 25-34 anni, e di genere misto in proporzioni quasi uguali. La maggior parte delle persone presenta il diploma di scuola superiore come titolo di studio più alto e per quanto riguarda l'occupazione si dividono equamente tra studenti, lavoratori indipendenti e dipendenti. Per quanto riguarda il rapporto con realtà virtuale e piattaforme di social VR, la maggior parte dei partecipanti riferisce di frequentare molto raramente questi ambienti e di avere poca esperienza con i sistemi di moderazione.

Procedura sperimentale

La procedura sperimentale è stata suddivisa in quattro fasi organizzate come segue:

- **Compilazione prima parte questionario e introduzione all'applicazione**: in questa fase iniziale, ai partecipanti è stato spiegato in maniera generica quale fosse lo scopo del test ed è stato chiesto loro di compilare la prima parte del questionario relativa alle informazioni generali, dopodiché sono stati introdotti all'ambiente virtuale, con particolare attenzione alla navigazione e agli strumenti di moderazione a loro disposizione. Per cercare di rendere l'esperienza il più sovrapponibile possibile con la realtà, ai partecipanti è stato detto che gli avatar sarebbero stati pilotati da un collaboratore, quindi una persona in carne ed ossa, da remoto, che tra le varie possibilità avrebbe sentito i loro messaggi, senza avere a disposizione nessuna informazione riguardo la loro identità, e avrebbe scelto la risposta da dare tra un set limitato di opzioni a disposizione.
- **Fase 1**: finita la parte introduttiva, ha avuto inizio la prima fase del test, durante la quale i partecipanti hanno effettivamente iniziato ad usare l'applicazione e a navigare l'ambiente virtuale interagendo con i quattro bot, avendo a disposizione solo gli strumenti di moderazione, quindi senza che fosse possibile l'intervento del sistema di moderazione automatica. La

durata di questa fase è stata di 5 minuti totali e al termine è stato chiesto ai partecipanti di compilare la parte del questionario relativo alla loro esperienza durante la fase 1.

- **Fase 2**: una volta terminata la compilazione del questionario, ha avuto inizio la seconda fase del test, la cui unica differenza rispetto alla fase precedente consiste nella presenza, oltre che degli strumenti di moderazione disponibili già nella fase 1, del sistema di moderazione automatica attivo. Ai partecipanti non è mai stato detto esplicitamente che sarebbe stato introdotto un sistema del genere fin dall'inizio della sperimentazione in modo da non influenzare il modo in cui avrebbero interagito con i bot. Anche questa fase ha avuto una durata di 5 minuti e al termine è stato chiesto ai partecipanti di compilare l'ultima parte del questionario, relativa alla loro esperienza durante la fase 2.

Misurazioni

Questa si focalizzerà su quali misure sono state effettuate durante lo svolgimento dei test e dopo attraverso l'analisi dei risultati dei questionari.

Misure obiettive

Per quanto riguarda le azioni intraprese dagli utenti, per ogni singola azione, sono stati registrati durante tutta la sperimentazione:

- Il tipo di azione intrapresa, quindi se il partecipante avesse parlato, mutato, bloccato o segnalato un bot, o avesse attivato la bolla personale.
- Il target dell'interazione nel caso in cui si trattasse di un'interazione con un bot.
- Lo stato a cui ha portato l'azione nel caso riguardasse la bolla personale.

Questo ha permesso di conseguenza di avere a disposizione per ogni strumento di moderazione il numero di volte che è stato utilizzato da ogni partecipante durante entrambe le fasi.

Per quanto riguarda il sistema di moderazione automatico, durante la seconda fase sono stati misurati:

- Numero di interventi per bot e numero di interventi nei confronti dell'utente.
- Tempo necessario per il sistema per rilevare, analizzare e reagire ai messaggi inappropriati.

Valutazioni soggettive

Sono stati valutati, attraverso opportune domande presenti nel questionario, alcuni aspetti comuni ad entrambi gli scenari che hanno permesso di effettuare successivamente un'analisi comparativa dei diversi sistemi di moderazione, in particolare:

- Stato emotivo complessivo dei partecipanti durante entrambe le fasi.
- Efficacia di ogni strumento e del sistema di moderazione automatica nel prevenire interazioni indesiderate o spiacevoli.
- Percezione di sicurezza e protezione nell'ambiente di social VR grazie alla disponibilità e all'utilizzo di questi strumenti e all'intervento del sistema di moderazione automatica.
- Trasparenza delle regole e delle azioni associate ad ogni strumento e al sistema di moderazione automatica (es. chiarezza sulle conseguenze delle segnalazioni o sulle regole in base al quale il sistema automatico bloccava i messaggi).
- Tempestività della risposta di ogni strumento e del sistema di moderazione automatica.
- Effetto avuto sulla comunicazione e le azioni e in generale sull'esperienza dei partecipanti dai due differenti scenari di moderazione.
- Impatto sulla possibilità degli utenti di esprimersi e di interagire liberamente con i bot.
- Percezione della protezione della propria privacy in entrambi gli scenari.
- Soddisfazione generale rispetto ai due diversi scenari di moderazione.

Per quanto riguarda gli strumenti di moderazione sono state effettuate alcune valutazioni esclusive:

- Facilità d’uso di ogni strumento.

Anche per il sistema di moderazione sono state effettuate delle valutazioni esclusive:

- Capacità del sistema di comprendere il significato e il contesto dei messaggi.
- Imparzialità e assenza di discriminazioni da parte del sistema.
- Qualità e accuratezza della moderazione automatica.
- Discrepanze col sistema percepite dai partecipanti su cosa sia considerabile inappropriato e non accettabile e cosa non lo è.

Risultati

Di seguito vengono riportati i risultati più significativi delle sperimentazioni. Attraverso un'analisi dettagliata delle misurazioni oggettive e delle valutazioni soggettive raccolte tramite questionari, esploreremo le dinamiche dell'esperienza dell'utente in entrambi gli scenari di moderazione. I dati qui presentati offriranno uno sguardo approfondito su come gli utenti hanno percepito e interagito con gli strumenti di moderazione.

I dati raccolti durante le sperimentazioni mostrano come la frequenza di utilizzo degli strumenti di moderazione sia diminuita, ma non cessata, con l’intervento del

sistema di moderazione automatica, e che le azioni più intraprese sono state in entrambi i casi mutare e bloccare gli utenti.

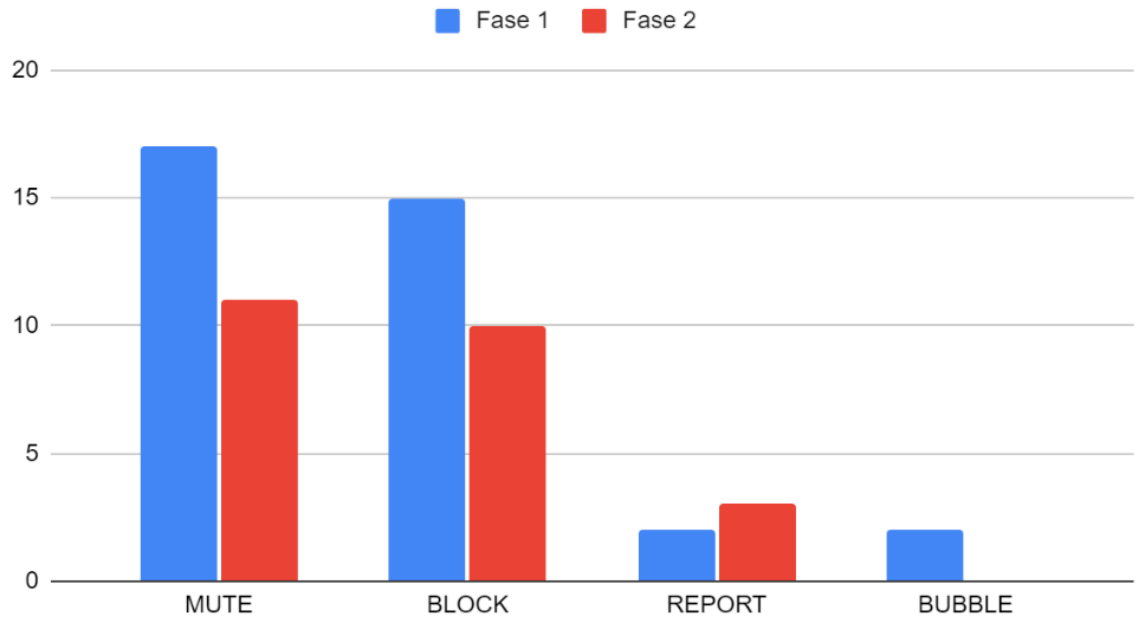


Figura 9 Utilizzo degli strumenti di moderazione

Di seguito uno sguardo più approfondito su quanto siano stati utilizzati i diversi strumenti di moderazione in base a quale fosse il bot target delle interazioni durante le due fasi del test.

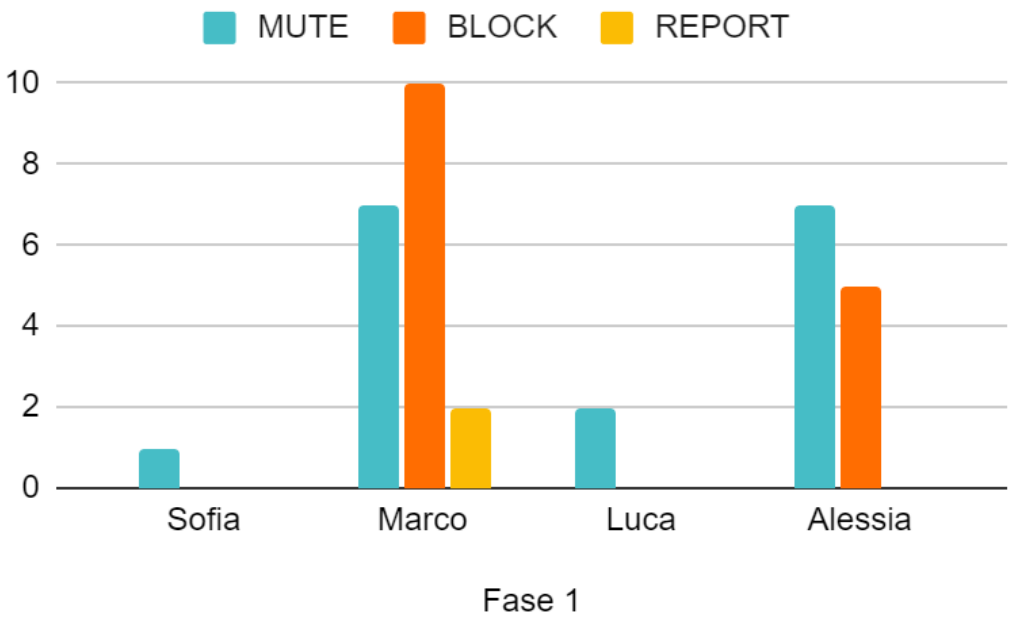


Figura 11 Volte in è stato utilizzato ogni strumento di moderazione per ogni bot (fase 1)

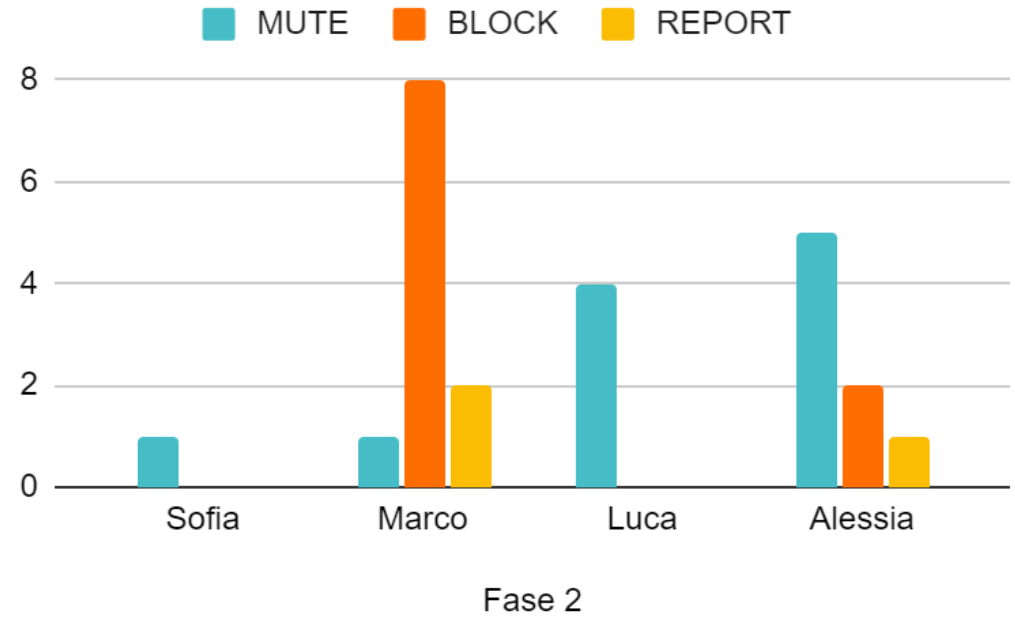


Figura 10 Volte in è stato utilizzato ogni strumento di moderazione per ogni bot (fase 2)

Possiamo osservare come i bot più colpiti dalle azioni dei partecipanti siano stati i due che assumevano i comportamenti più invadenti e inappropriati e come sia

variata l'interazione con loro tra una fase l'altra, per esempio nel caso di Marco i partecipanti, con la presenza del sistema di moderazione automatica, hanno sentito meno la necessità di mutarlo, ma hanno sentito comunque la necessità di bloccarlo dal momento che una delle peculiarità di Marco era il fatto che dopo poche interazioni con l'utente iniziava ad inseguirlo e ad invadere il suo spazio personale.

Per quanto il sistema di moderazione automatica, è stato possibile misurare una media della frequenza degli interventi per ogni bot e nei confronti dei partecipanti:

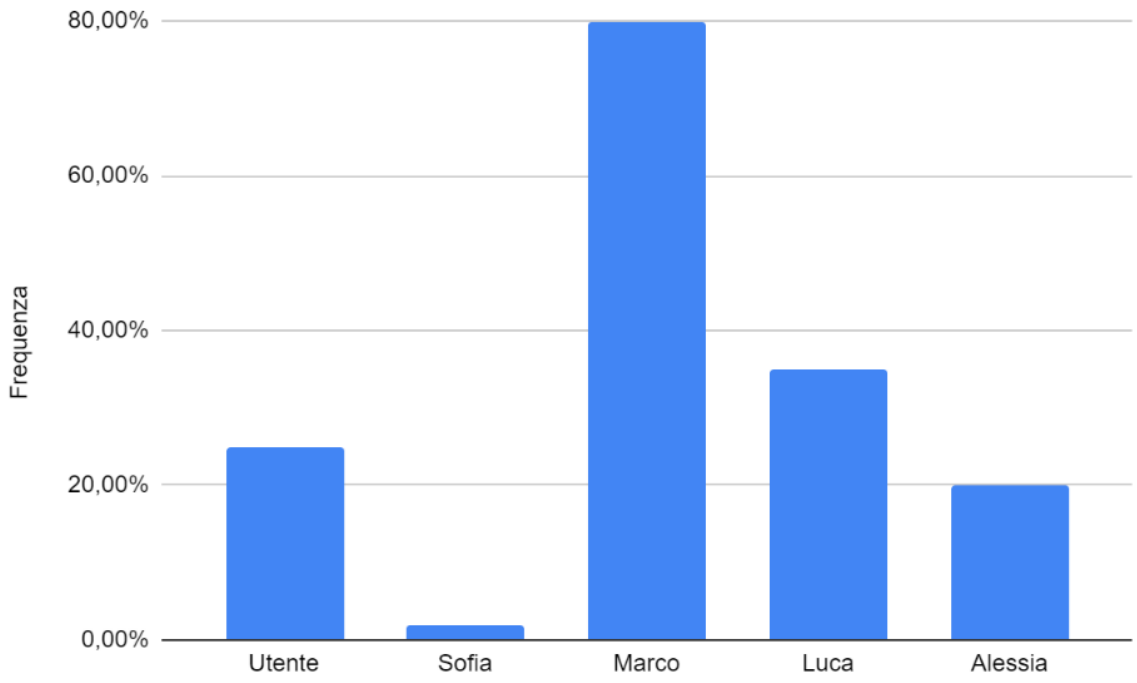


Figura 12 Frequenza con cui è intervenuto il sistema di moderazione automatica

È stato misurato anche il tempo di riposta medio del sistema, che comprende il tempo necessario per effettuare la conversione dell'audio in testo e la classificazione del testo e in media corrisponde a 315 ms.

Analizziamo adesso i risultati dei questionari somministrati ai partecipanti. Mentre per la maggior parte dei criteri valutati non sono emerse particolari differenze tra uno scenario e l'altro, con una tendenza degli strumenti di moderazione ad ottenere valutazioni leggermente più alte rispetto al sistema di moderazione automatica, alcuni punti hanno mostrato degli interessanti sbilanciamenti da una parte o dall'altra, a partire dall'efficacia dei due sistemi di moderazione.

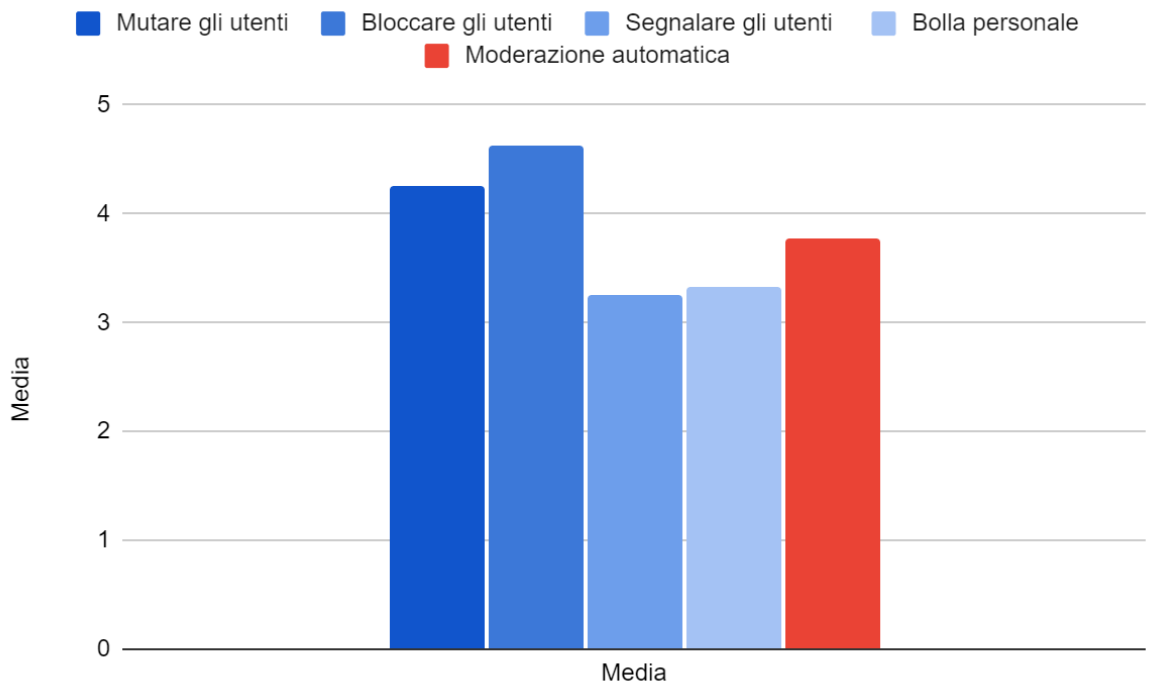


Figura 13 Efficacia media percepita dei vari sistemi di moderazione.

Come possiamo vedere, la bolla personale dà prova di risultare più efficace per prevenire invasioni degli spazi personali e in generale interazioni indesiderate di tipo fisico, mentre la segnalazione, con i tempi di attesa elevati che è necessario affrontare prima di vedere effettivamente qualche risultato, non è riuscita a soddisfare le esigenze degli utenti. I partecipanti, però, sembrano considerare più efficace poter mutare e bloccare gli altri utenti, rispetto ad affidarsi al sistema di moderazione automatica. Questo potrebbe suggerire che, sebbene il sistema di moderazione automatica possa ridurre il carico di lavoro della moderazione umana, non può sostituirla completamente.

Inoltre, in presenza del sistema di moderazione automatica, i partecipanti hanno sentito una maggiore limitazione della loro possibilità di esprimersi e di interagire con i bot, questo perché molti messaggi sono stati bloccati e non sempre con un'apparente motivazione valida, infatti spesso il sistema risultava sensibile anche a forme di cosiddetta "mock impoliteness", ovvero un tipo di linguaggio o comportamento che, sebbene appaia scortese o maleducato in superficie, è in realtà utilizzato in modo scherzoso o giocoso e non intende ferire o offendere veramente.

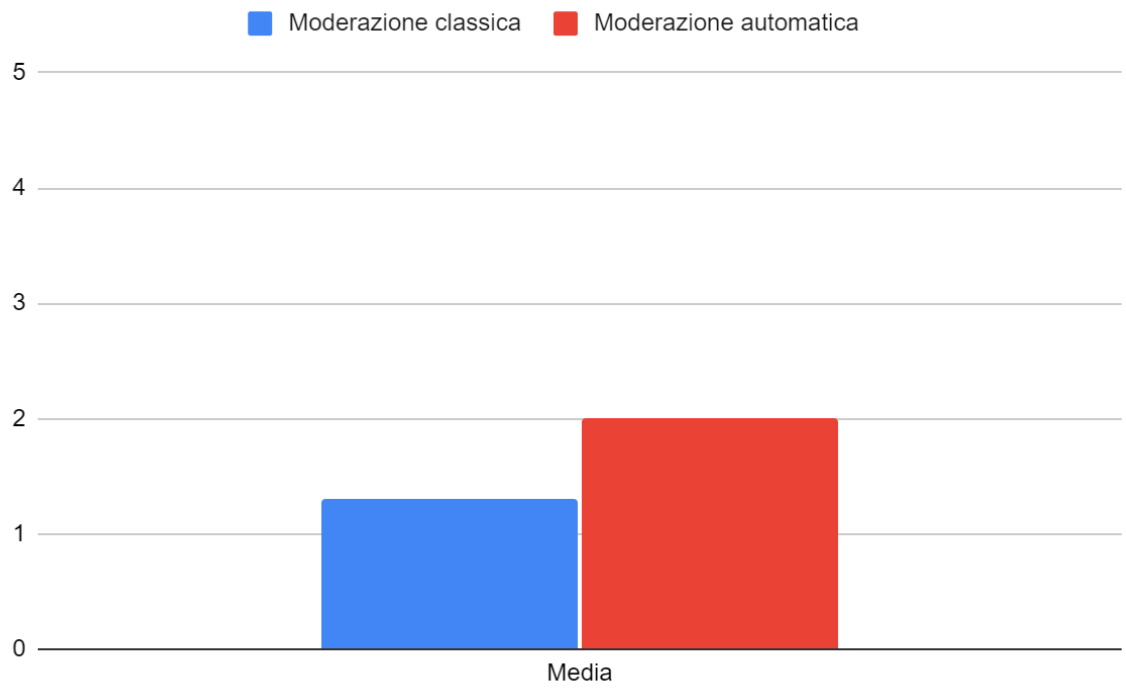


Figura 14 Limitazione media percepita con i due sistemi di moderazione

Un altro punto cruciale in cui sono emerse delle sostanziali differenze tra i due scenari è quello della privacy; infatti, in presenza del sistema di moderazione automatica la maggior parte delle persone si sono sentite poco protette nella loro privacy essendo consapevoli del fatto che quello che dicevano veniva continuamente ascoltato e processato automaticamente, ma senza sapere né come, né dove e né da chi.

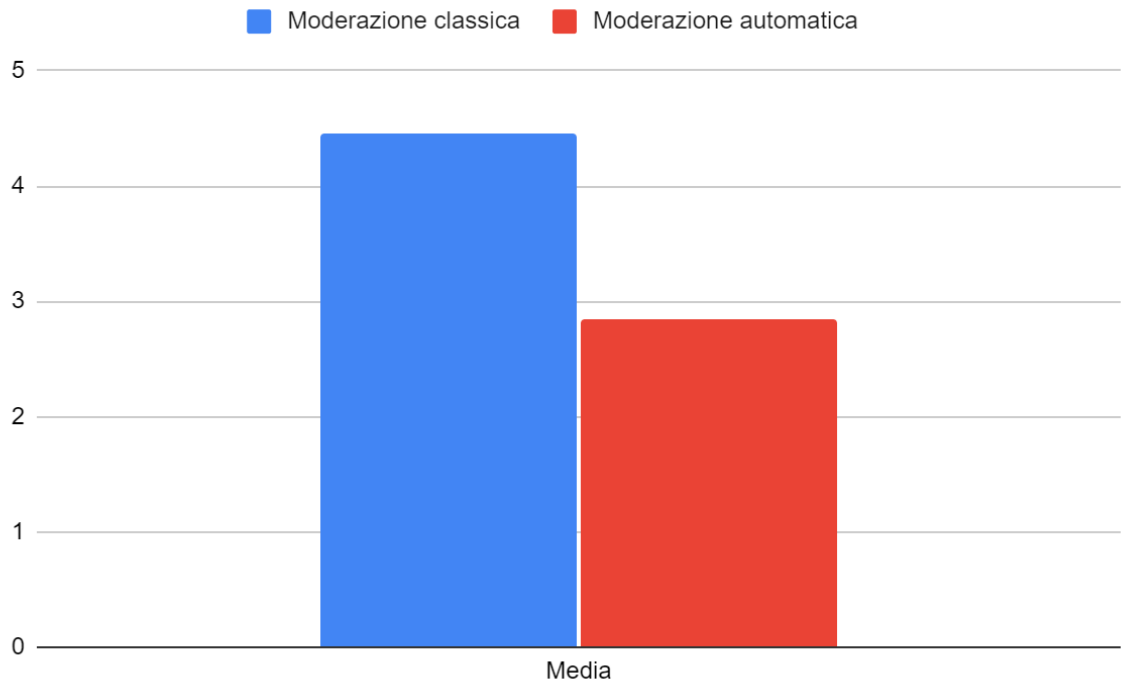


Figura 15 Protezione della privacy media percepita con i due sistemi di moderazione

Questo ha comportato un leggero vantaggio degli strumenti di moderazione rispetto al sistema quando è stato chiesto hai partecipanti di valutare l'impatto dei due sistemi sulla loro esperienza e la propria soddisfazione complessiva.

Un altro dato interessante è il fatto che, secondo le risposte dei partecipanti, in presenza del sistema di moderazione automatica hanno avvertito maggiormente una sorta di legittimazione a comportarsi in modo inappropriato rispetto a quando avevano a disposizione solo gli strumenti di moderazione più classici.

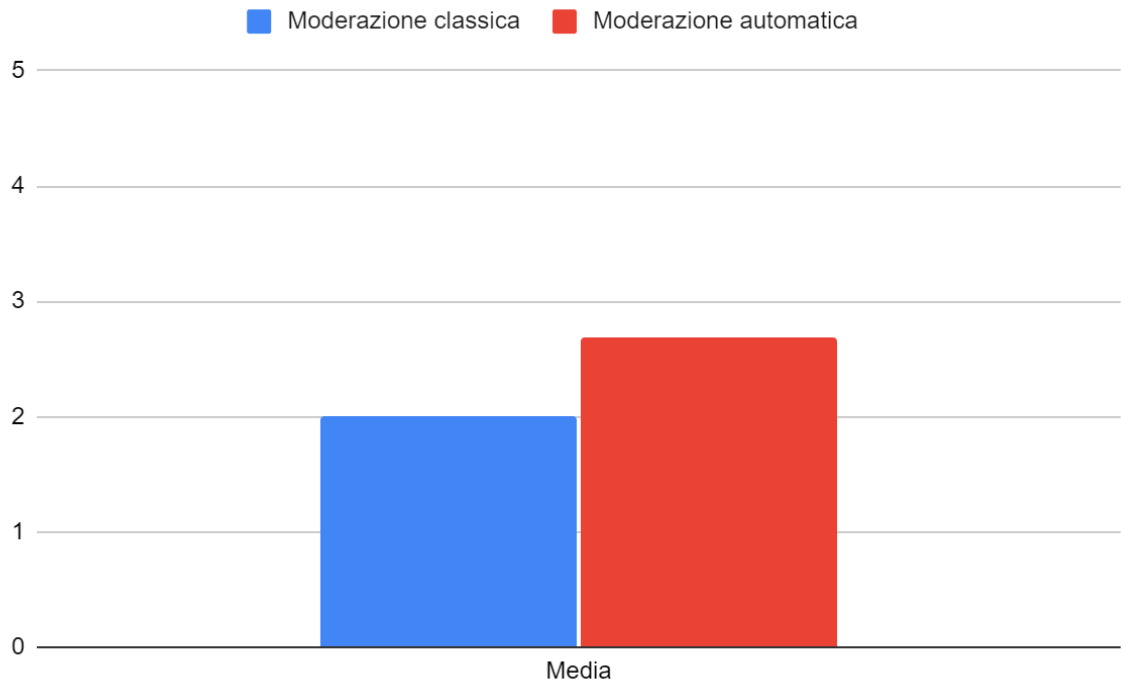


Figura 16 Legittimazione a comportarsi in modo inappropriato media percepita

Questo effetto potrebbe essere dovuto alla consapevolezza che i diretti interessati non avrebbero sentito le loro ingiurie e al fatto che durante la simulazione i partecipanti non sono stati puniti in nessun modo dal sistema; infatti, le uniche ritorsioni che potevano ritrovarsi ad affrontare durante tutta la sperimentazione erano le azioni degli altri bot. Sarebbe interessante in futuro esplorare questa dinamica in un contesto più realistico.

Conclusioni e considerazioni future

Tenendo in considerazione che questo studio è stato effettuato su un campione di persone piccolo e non molto variegato e che la simulazione ha cercato di riprodurre più aspetti possibili dell'esperienza di un utente in una ambienti virtuali quali sono le piattaforme di social VR, portandosi però dietro una serie di limitazioni, come la presenza di soli altri quattro utenti all'interno della stanza e il fatto che questi utenti fossero impersonificati da dei bot e non da delle persone reali, penalizzando così la naturalezza e l'immersività dell'esperienza, non sembra che gli utenti abbiano tratto particolari vantaggi dalla presenza del sistema di moderazione automatica all'interno dell'ambiente e viene spontaneo chiedersi se, considerando gli svantaggi introdotti, valga davvero la pena di implementare una moderazione proattiva basata sull'intelligenza artificiale e se gli strumenti di moderazione non siano sufficienti per permettere agli utenti di proteggersi.

Sarebbe utile valutare l'impatto sull'esperienza in presenza di molti più utenti nello stesso ambiente, in questo caso le cose potrebbero cambiare perché gestire un numero elevato di utenti avendo a disposizione solo strumenti di moderazione più classici potrebbe risultare difficile. Un altro contesto in cui il sistema di moderazione automatica potrebbe risultare più utile e che non è stato testato in questo studio a causa della natura delle interazioni con alcuni dei bot, è quello di ambienti in cui coabitano minorenni, in particolare ragazzi di età tra i 13 e i 17 anni non essendo

possibile accedere a queste piattaforme con un età inferiore, e adulti perché, come accennato all'inizio di questo documento, non sono rari i casi in cui avvengono molestie da parte di adulti nei confronti di minorenni e quest'ultimi potrebbero non essere abbastanza consapevoli o avere la prontezza di usare gli strumenti di moderazione autonomamente e in maniera efficiente.

In questo caso potrebbero essere presi in considerazione alcuni possibili miglioramenti a cui sottoporre il sistema di moderazione automatica, tra cui:

- Migliorare l'accuratezza del sistema nel distinguere tra comportamenti veramente offensivi e comportamenti scherzosi o giocosi e la capacità di comprendere il contesto della conversazione.
- Migliorare i tempi di risposta del sistema di moderazione automatica per rendere la moderazione real-time e quindi aumentarne l'azione preventiva.
- Esplorare maggiormente come tutelare la privacy degli utenti mentre si utilizza un sistema di moderazione automatica ed essere il più trasparenti possibile riguardo il funzionamento del sistema.

Potrebbe essere interessante esplorare come gli strumenti di moderazione potrebbero essere personalizzati in base alle preferenze individuali degli utenti. Per esempio, l'utilizzo di filtri associati alla moderazione automatica potrebbe risultare particolarmente adatto alla personalizzazione dell'esperienza utente. Per esempio,

se non si vuole impedire il dibattito su un argomento controverso come quello della guerra, e ricordiamo che la censura del dibattito su argomenti ritenuti controversi è un problema ancora presente in ambienti in cui viene effettuata moderazione automatica più frequentemente, come le piattaforme social tradizionali, gli utenti potrebbero avere la possibilità di settare dei filtri che permettano alla piattaforma di avvisarli quando stanno per accedere ad un ambiente o visualizzare un contenuto in cui il dibattito è aperto o evitare direttamente che possano accedervi, per esempio nel caso di utenti minorenni, e allo stesso tempo garantire a questi utenti di poter frequentare ambienti in cui questo dibattito sarebbe soggetto a moderazione.

Dei tool che hanno una funzione simile sono il Mute Assist implementato recentemente da Meta Horizon Worlds, che permette di scegliere se bloccare o no i messaggi degli utenti classificati come inopportuni e con quale restrittività classificare questi messaggi, o Tune, un'estensione sperimentale di Google Chrome che permette di impostare dei filtri per nascondere commenti che vengono classificati da Tune secondo diverse categorie sovrapponibili a quelle di Perspective API e che si può scegliere se bloccare o no.

Fondamentale per il futuro sarà condurre ulteriori sperimentazioni in contesti più realistici per comprendere meglio come le dinamiche di moderazione influenzano il comportamento degli utenti e approfondire lo studio sull'impatto psicologico

della presenza del sistema di moderazione automatica e sulla percezione della legittimità del comportamento degli utenti.

Riferimenti

[1] D. Castro, «Content Moderation in Multi-User Immersive Experiences: AR/VR and the Future of Online Speech,» 2022.

[2] J. Belamire, «My First Virtual Reality Groping,» 20 Ottobre 2016. [Online]. Available: https://medium.com/athena-talks/my-first-virtual-reality-sexual-assault-2330410b62ee#.8lcy2o2bh.

[3] K. Shriram e R. Schwartz, «All Are Welcome: Using VR Ethnography to Explore Harassment Behavior in Immersive Social Virtual Reality,» 2017.

[4] J. Outlaw, «Virtual harassment: The social experience of 600+ regular virtual reality (VR),» 2018.

[5] L. Blackwell, N. Ellison, N. Elliott-Deflo e R. Schwartz, «Harassment in Social Virtual Reality: Challenges for Platform Governance,» *Proceedings of the ACM on Human-Computer Interaction,* 2019.

[6] CCDH, «Facebook's Metaverse,» 2021. [Online]. Available: https://counterhate.com/research/facebooks-metaverse/.

[7] N. Sabri, B. Chen, A. Teoh, S. P. Dow, K. Vaccaro e M. Elsherief, «Challenges of Moderating Social Virtual Reality,» 2023.

[8] G. Freeman, S. Zamanifard, D. Maloney e D. Acena, «Disturbing the Peace: Experiencing and Mitigating Emerging Harassment in Social Virtual Reality,» *Proceedings of the ACM on Human-Computer Interaction,* 2022.

[9] B. K. Wiederhold, «Sexual Harassment in the Metaverse. Cyberpsychology, Behavior, and Social Networking.,» 2022.

[10] J. A. Jiang, C. Kiene, S. Middler, J. R. Brubaker e C. Fiesler, «Moderation Challenges in Voice-based Online Communities on Discord,» 2019.

[11] D. Maloney, G. Freeman e A. Robb, «A Virtual Space for All: Exploring Children's Experience in Social Virtual Reality,» 2020.

[12] M. Steiger, T. J. Bharucha, S. Venkatagiri, M. J. Riedl e M. Lease, «The Psychological Well-Being of Content Moderators: The Emotional Labor of Commercial Moderation and Avenues for Improving Support,» 2021.

[13] R. Gorwa, R. Binns e C. Katzenbach, «Algorithmic content moderation: Technical and political challenges in the automation of platform governance.,» 2020.

[14] «Global Internet Forum to Counter Terrorism,» [Online]. Available: https://gifct.org/.

[15] «PhotoDNA,» [Online]. Available: https://www.microsoft.com/photodna.

[16] T. Dias Oliva, D. Antonialli e A. Gomes, «Fighting Hate Speech, Silencing Drag Queens? Artificial Intelligence in Content Moderation and Risks to LGBTQ Voices Online.,» 2020.

[17] «Rec Room,» [Online]. Available: https://recroom.com/.

[18] «VRChat,» [Online]. Available: https://hello.vrchat.com/.

[19] «Meta Horizon Worlds,» [Online]. Available: https://horizon.meta.com/.

[20] «OpenAI Moderation,» [Online]. Available: https://platform.openai.com/docs/guides/moderation.

[21] «Perspective API,» [Online]. Available: https://perspectiveapi.com/.

[22] «Azure AI Content Moderator,» [Online]. Available: https://learn.microsoft.com/en-us/azure/ai-services/content-moderator/.

[23] Jigsaw, «https://developers.perspectiveapi.com/,» [Online].

[24] «Microsoft Azure Speech Service,» [Online]. Available: https://learn.microsoft.com/en-us/azure/ai-services/speech-service/.

[25] N. Petherbridge, «RiveScript,» 2023. [Online]. Available: https://www.rivescript.com/.

[26] F. Ávila, «rivescript-csharp,» GitHub, 2021. [Online]. Available: https://github.com/fabioravila/rivescript-csharp.

[27] «Unity,» [Online]. Available: https://unity.com/.

[28] «VRoid Studio,» [Online]. Available: https://vroid.com/en/studio.

[29] «Blender,» [Online]. Available: https://www.blender.org/.

[30] «VRM Blender Add-on,» [Online]. Available: https://github.com/saturday06/VRM-Addon-for-Blender.

[31] «Cats Blender Add-On,» [Online]. Available: https://github.com/absolute-quantum/cats-blender-plugin.

[32] «Material Combiner Blender Add-on,» [Online]. Available: https://github.com/Grim-es/material-combiner-addon.

[33] «Mixamo,» [Online]. Available: https://www.mixamo.com.

[34] «Rec Room Code of Conduct,» [Online]. Available: https://recroom.zendesk.com/hc/en-us/articles/4419890420887-Rec-Room-Code-of-Conduct.

[35] M. Horta Ribeiro, J. Cheng e R. West, «Automated Content Moderation Increases Adherence to Community Guidelines,» 2023.