

WireCrossed

**Developing Community Media to Mitigate
the Impact of fake news**

**Modulo 2 - Modelli di
moderazione dei
contenuti e loro
applicabilità**

The image shows a slide layout. On the left is a solid black vertical bar. On the right is a white rectangular area containing the 'WireCrossed' logo, which consists of the words 'Wire' and 'Crossed' in a stylized font with lines radiating from the 'C'. Below the logo is the subtitle 'Developing Community Media to Mitigate the Impact of fake news'. At the bottom of the white area is the main title 'Modulo 2 - Modelli di moderazione dei contenuti e loro applicabilità' in a bold, dark red font.



Ciao !



Panoramica

La moderazione dei contenuti è un processo che riceve un'attenzione limitata e spesso passa inosservato nella nostra esperienza su Internet. Allo stesso tempo essa è considerata una necessità per garantire agli utenti una navigazione ed un'esperienza sicura.

Attraverso il funzionamento di questo workshop ti verrà fornita una panoramica del termine, nonché dei problemi e degli approcci che comporta.

Argomenti

- Cos'è la moderazione dei contenuti e qual è la sua importanza
- I vari modelli di moderazione dei contenuti
- Strumenti tecnici e approcci nella moderazione dei contenuti
- Black-list e filtri di pubblicazione
- Le pratiche di moderazione dei contenuti adottate dai social media

Il workshop cercherà di trattare i seguenti argomenti:

- Cos'è la moderazione dei contenuti e qual è la sua importanza
- I vari modelli di moderazione dei contenuti
- Strumenti tecnici e approcci nella moderazione dei contenuti
- Black-list e filtri di pubblicazione
- Le pratiche di moderazione dei contenuti adottate dai social media

Obiettivi formativi

- Comprendere cos'è la moderazione dei contenuti e la sua necessità
- Imparare a conoscere i diversi tipi di moderazione e i loro vantaggi
- Scoprire come creare una "black-list" e come applicare filtri di post pubblicazione
- Apprendere le pratiche utilizzate dai fornitori di servizi di social media per identificare le anomalie nella diffusione

Servendosi della formazione che verrà loro offerta e del materiale di apprendimento autogestito del modulo, gli studenti:

Impareranno a conoscere i diversi tipi di moderazione e i loro vantaggi

Acquisiranno conoscenze di base sui vari strumenti tecnici

Saranno in grado di riconoscere le caratteristiche dei contenuti social intenzionalmente negativi e provocatori

Scopriranno come creare una "black-list" e come applicare filtri di post pubblicazione

Apprenderanno le pratiche utilizzate dai fornitori di servizi di social media per identificare le anomalie nella diffusione

Presentazione della teoria

**Moderazione dei contenuti e
sua importanza**



Introduzione

- L'espansione delle tecnologie digitali, con la diffusione di Internet, ha portato ad una "democratizzazione dell'informazione"
- Oggi, per tenersi informati, i cittadini non si affidano solo ai media tradizionali (TV, radio e giornali)
- Essi hanno accesso ad un'abbondanza di fonti diverse e possono inoltre condividere i propri pensieri e le proprie esperienze

Servendosi della formazione che verrà loro offerta e del materiale di apprendimento autogestito del modulo, gli studenti:

Impareranno a conoscere i diversi tipi di moderazione e i loro vantaggi

Acquisiranno conoscenze di base su vari strumenti tecnici

Saranno in grado di riconoscere le caratteristiche dei contenuti social intenzionalmente negativi e provocatori

Scopriranno come creare una «blacklist» e come applicare filtri di post pubblicazione

Apprenderanno le pratiche utilizzate dai fornitori di servizi di social media per identificare le anomalie nella diffusione

Introduzione

- Nonostante l'ottimismo iniziale, oggi questo sviluppo non è più considerato del tutto positivo
- Internet è diventato uno spazio in cui è stato condiviso anche materiale minaccioso e dannoso
- È necessario adottare delle misure per garantire che il contenuto condiviso e accessibile dagli utenti non sia dannoso

Questo sviluppo tecnologico è stato accolto con grande entusiasmo e speranza. Molti analisti hanno percepito questa evoluzione come l'alba di una grande epoca in cui gli individui saranno più informati e il discorso pubblico beneficerà in modo esponenziale dalla partecipazione attiva di tutti i cittadini e dalla presentazione di tutti i punti di vista. Tuttavia, questo ottimismo è stato presto infranto, in quanto la sfera digitale ha anche offerto terreno fertile per la condivisione e per il fiorire di materiale minaccioso e spesso pieno di odio, nonché per la diffusione incontrollata e spesso deliberata di informazioni false.

Di conseguenza è della massima importanza che ogni forma di media introduca e metta in atto procedure di moderazione dei contenuti, per garantire che il contenuto che viene presentato o ospitato non sia dannoso. Sono necessari passaggi concreti per garantire che i contenuti potenzialmente dannosi sotto forma di materiale pedopornografico, contenuti violenti ed estremi, incitamento all'odio, contenuti grafici, contenuti sessuali, materiale crudele e insensibile e contenuti spam siano bloccati (Ofcom: 2019). A tal fine il presente modulo desidera introdurre specifici modelli di moderazione dei media, con risorse aggiuntive, nel tentativo di supportare gli studenti interessati allo sviluppo dei media comunitari nello svolgimento efficace di tali processi.

Cos'è la moderazione dei contenuti?

- Il Cambridge Dictionary definisce la moderazione come la qualità del fare qualcosa entro limiti ragionevoli.
- Nel contesto dei media, questo principio si riferisce alla definizione di regolamenti e limiti per quanto riguarda le informazioni e il materiale da condividere o ospitare
- Si riferisce al monitoraggio del contenuto caricato e al blocco o alla rimozione di qualsiasi contenuto ritenuto non accettabile

Il Cambridge Dictionary definisce la moderazione come la qualità del fare qualcosa entro limiti ragionevoli. Nel contesto dei contenuti dei media, questo principio si riferisce alla definizione di regolamenti e limiti per quanto riguarda le informazioni e il materiale da condividere o ospitare su una piattaforma, che sia un organo di stampa, un sito social, un sito web. In altre parole, si riferisce al monitoraggio del contenuto che è o è già stato caricato e non consente o rimuove qualsiasi contenuto che non sia accettabile dall'insieme di regole in atto sulla piattaforma in questione (Grimes-Viort: 2010). "A seconda dei requisiti specifici di ogni sito, la moderazione può essere intrapresa in misura maggiore o minore. Alcuni forum sono orgogliosi della libertà di parola, mentre altri, come i siti di social networking, devono trovare un difficile equilibrio tra facilità d'uso e protezione dei loro utenti più giovani "(Smith: 2019).

Cos'è la moderazione dei contenuti?

- La stragrande maggioranza delle piattaforme di social media e dei media (digitali o tradizionali) ha specifiche linee guida di moderazione che determinano in una certa misura quale tipo di contenuto è accettabile per la diffusione
- Rispettando queste regole e impiegando varie tecniche e approcci di moderazione dei contenuti, essi monitorano e valutano il materiale condiviso con o tramite loro

La maggior parte dei social media e dei media popolari (digitali o tradizionali) hanno linee guida di moderazione specifiche che determinano in una certa misura quale tipo di contenuto è accettabile per la diffusione attraverso la loro piattaforma. Rispettando queste regole e impiegando varie tecniche e approcci di moderazione dei contenuti, essi monitorano e valutano il materiale condiviso per garantire che tutti i tipi di contenuti offensivi o discutibili come video porno, immagini esplicite o immagini non adatti a tutte le fasce d'età non siano pubblicati o rimossi (Cogito Tech: 2020).

Perché è necessaria la moderazione dei contenuti?

- È fondamentale garantire che il pubblico / gli utenti di un media o di una piattaforma non siano esposti a nulla di discutibile o dannoso
- Ospitare o condividere comportamenti inappropriati, dannosi o illegali non solo causerà problemi alla piattaforma / ai media, ma potrebbe avere gravi implicazioni per il loro pubblico, in particolare per i giovani
- Pertanto è della massima importanza che i media e le piattaforme in questione siano vigili

La moderazione delle immagini, dei video e del testo sono fondamentali per garantire che il pubblico / gli utenti di un media o di una piattaforma non siano esposti a nulla di discutibile o dannoso (Schomer: 2019). Ospitare o condividere comportamenti inappropriati, dannosi o illegali non solo causerà problemi alla piattaforma / ai media, ma potrebbe avere gravi implicazioni per il loro pubblico, in particolare per i giovani che utilizzano Internet e prevalentemente i social media per accedere a informazioni e notizie così come per le interazioni sociali. Pertanto è della massima importanza che i media e le piattaforme in questione siano vigili e che investano nel monitoraggio e nella moderazione dei loro contenuti fruibili (puremoderation: 2020).

Moderazione dei contenuti in prospettiva





ATTIVITA' Nr. 1
Le minacce con cui i moderatori hanno a che fare

Tramite i vostri telefoni inviate a Mentimeter i tipi di minacce a cui state pensando.

Teoria - Presentazione: modelli di moderazione dei contenuti



Esistono diversi approcci o modelli di moderazione dei contenuti che una piattaforma può seguire

Questi modelli in una certa misura non si escludono a vicenda, quindi si potrebbe scegliere di mettere in atto più approcci.

In definitiva si deve decidere quale è il più efficace e quale funziona meglio a questo scopo

Quando si parla di moderazione dei contenuti ci sono diversi approcci o modelli che una piattaforma può impiegare, a seconda dei propri principi, modus operandi, obiettivi e processi di lavoro. Questi modelli in una certa misura non si escludono a vicenda, quindi si potrebbe scegliere di metterne in atto diversi. In definitiva si dovrebbe decidere l'approccio da adottare con l'obiettivo di salvaguardare il miglior risultato possibile di monitoraggio e filtraggio, quello che proteggerà il pubblico da contenuti dannosi e sconcertanti, oltre a mantenere un certo senso di ordine nel suo funzionamento (Cogito Tech: 2019).

Comprendere i diversi tipi di moderazione dei contenuti, insieme ai loro punti di forza e di debolezza, può aiutare un moderatore / gestore dei media a prendere la decisione giusta, quella che opererà al meglio per il suo proposito e il suo funzionamento.



Pre-moderazione

- Il modello di pre-moderazione implica l'impostazione di un processo in cui tutto il contenuto richiede l'approvazione prima di essere pubblicato.
- Qualsiasi contenuto inviato a una piattaforma viene messo in coda per essere esaminato e approvato da un moderatore, prima che diventi visibile al pubblico degli utenti.
- Assicura che qualsiasi contenuto che non rispetta le regole di un sito / piattaforma sia tenuto fuori dalle sue sezioni pubbliche visibili.
- Offre un alto controllo al manager-moderatore.

Fonte dell'immagine: imgaga.com/webinar-content-moderation

Come indica il nome, il modello di pre-moderazione prevede l'impostazione di un processo in cui tutto il contenuto richiede l'approvazione prima di essere pubblicato. In termini pratici, qualsiasi contenuto inviato a una piattaforma viene messo in coda per essere esaminato e approvato da un moderatore, prima che diventi visibile al pubblico degli utenti. La pre-moderazione aiuta a garantire che (nelle mani di un buon moderatore) qualsiasi contenuto che potrebbe essere dannoso, inappropriato o non conforme alle regole di un sito / piattaforma sia tenuto fuori dalle sue sezioni pubbliche visibili. In questo modo offre un alto controllo al manager-moderatore.

Svantaggi della pre-moderazione

- Potrebbe causare reclami da parte dell'individuo che invia il contenuto
- Può portare a ritardi che potrebbero avere un effetto negativo sui contenuti che sono colloquiali e quindi sensibili al fattore tempo
- Comporta un grande sforzo che potrebbe diventare ingestibile

Tuttavia la Pre-moderazione ha i suoi svantaggi, in quanto può causare lamentele da parte dell'individuo che invia i contenuti o può portare a ritardi nel caricamento che possono avere un effetto negativo sui contenuti, i quali sono generalmente colloquiali e quindi sensibili al fattore tempo. Inoltre, questo è un processo che comporta un costo elevato. Se e quando una piattaforma andrà a crescere, e gli invii supereranno una certa soglia di contenuti generati dagli utenti, questi non saranno più gestibili dai moderatori (Grimes-Viort: 2010).

Post-moderazione

- Consentire la pubblicazione di contenuti / materiali su una piattaforma senza alcuna revisione iniziale e quindi senza valutazione
- Riduce al minimo i ritardi e rende più veloce il funzionamento di una piattaforma e le interazioni che si svolgono su / attraverso di essa
- Consente ai moderatori di garantire la sicurezza, mentre i problemi comportamentali e legali possono essere identificati e risolti in tempi ragionevoli



Source of the image: www.telusinternational.com/articles/mastering-content-moderation

Questo approccio prevede di consentire la pubblicazione di contenuti / materiali su una piattaforma senza alcuna revisione iniziale. Il moderatore può quindi valutarlo in seguito e, se necessario, rimuoverlo. Il vantaggio di questo tipo di moderazione è che riduce al minimo i ritardi e rende più veloce il funzionamento di una piattaforma e le interazioni che si svolgono su / attraverso di essa. "Le persone si aspettano un livello di immediatezza quando interagiscono sul web e la post-moderazione lo consente, permettendo anche ai moderatori di garantire che i problemi di sicurezza, comportamentali e legali possano essere identificati e risolti in tempi ragionevoli" (Grimes-Viort: 2010).

Svantaggi della Post-moderazione

- Potrebbe portare alcuni utenti ad imbattersi in contenuti dannosi o inappropriati.
- È un processo che può diventare oneroso man mano che la visibilità e la portata di una piattaforma crescono e di conseguenza i contenuti ad essa sottoposti aumentano.

Tuttavia, essa potrebbe portare alcuni utenti a imbattersi in contenuti dannosi o inappropriati. Inoltre, è un processo che può diventare oneroso man mano che la visibilità e la portata di una piattaforma crescono e di conseguenza aumentano i contenuti ad essa sottoposti.



Moderazione reattiva

- Un approccio che si basa sugli utenti per segnalare il contenuto, qualora essi lo ritengano inappropriato.
- Viene messa in atto una procedura di segnalazione tramite la quale gli utenti possono "contrassegnare" il contenuto e chiederne la rimozione.
- Può essere descritto come un tipo di moderazione dei contenuti generata dagli utenti stessi e indiscutibilmente mette potere e responsabilità nelle mani degli utenti-pubblico di una piattaforma

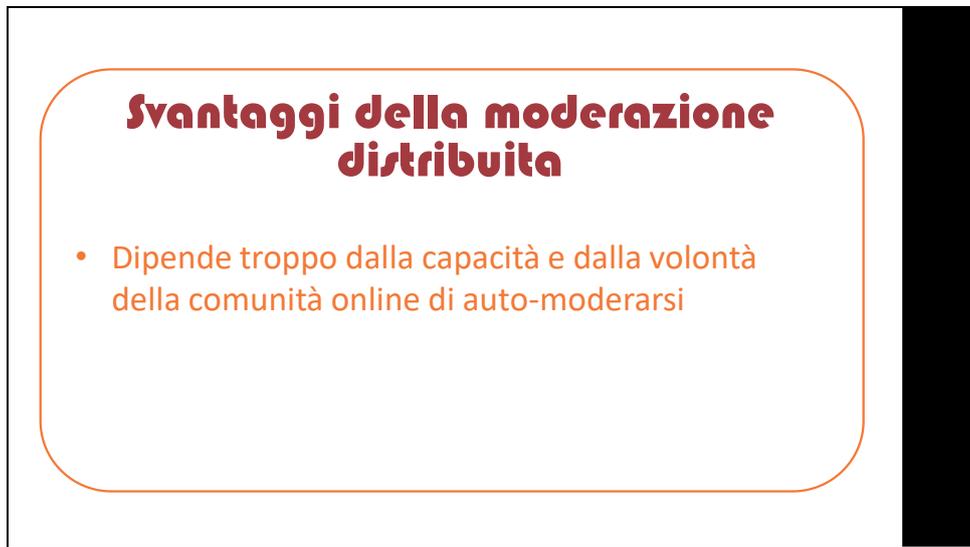
Source of the image: www.the-future-of-commerce.com/2019/10/17/social-media-and-customer-service-alignment/

La moderazione reattiva è un approccio che si basa sugli utenti per segnalare il contenuto qualora essi lo ritengano inappropriato. Viene messa in atto una procedura di segnalazione che fornisce agli utenti la possibilità di informare l'amministratore o il moderatore e chiedere la revisione e la rimozione di taluni contenuti. La moderazione reattiva è ampiamente utilizzata nei social media dove gli utenti possono contribuire attivamente alla moderazione del contenuto. Può essere descritto come un tipo di moderazione dei contenuti generata dagli utenti stessi e indiscutibilmente mette potere e responsabilità nelle mani degli utenti-pubblico di una piattaforma (Cogito Tech: 2019).

Svantaggi della moderazione reattiva

- Può portare ad un'eccessiva dipendenza dagli utenti e ad un controllo limitato.
- La moderazione dei contenuti diventa una questione di interpretazione e quindi di discussione tra gli utenti.

Il vantaggio principale di questo metodo di moderazione è che è proporzionale alla visibilità di una piattaforma e le consente perciò di eludere la responsabilità per i contenuti inappropriati caricati dagli utenti, purché questi rispondano alle richieste di revisione e rimozione di tale materiale. Tuttavia comporta un rischio, poiché porta a un'eccessiva dipendenza dagli utenti (Grimes-Viort: 2010).



Svantaggi della moderazione distribuita

- Dipende troppo dalla capacità e dalla volontà della comunità online di auto-moderarsi

È importante tenere in considerazione che questo approccio si basa sulla capacità e sulla volontà della comunità online di auto-moderarsi.

Moderazione automatizzata

- Implica l'introduzione e l'utilizzo di strumenti tecnici e processi di filtraggio automatizzati per controllare e rivedere i contenuti e gli invii
- Lo strumento più tipico utilizzato è il filtro delle parole, che esamina un testo alla ricerca di parole vietate, precedentemente definite, e le sostituisce o blocca del tutto il testo.
- Esistono strumenti simili, per la moderazione di immagini e video, i quali vanno a contrassegnare detti contenuti multimediali attraverso delle didascalie di inappropriatezza.

In aggiunta a tutti i suddetti sistemi di moderazione alimentati dall'uomo, c'è la moderazione automatizzata che è un'arma preziosa nell'arsenale del moderatore. Comprende l'introduzione e l'utilizzo di strumenti tecnici e di processi di filtraggio automatizzati per controllare e rivedere i contenuti e gli invii. Lo strumento più tipico utilizzato è il filtro delle parole, che esamina un testo alla ricerca di parole vietate, precedentemente definite, e le sostituisce o blocca del tutto il testo. Esistono strumenti simili, per la moderazione di immagini e video, i quali vanno a contrassegnare detti contenuti multimediali attraverso delle didascalie di inappropriatezza. (Grimes-Viort: 2010).

Svantaggi della moderazione automatizzata

- I sistemi automatizzati possono essere ingannati o non riuscire a rimuovere automaticamente materiale dannoso o minaccioso
- Possono mancare la percezione del contesto o l'approccio del materiale che stanno esaminando



BLACKLIST

ATTIVITA' Nr. 2

Creazione di una «black-list» e impostazione dei filtri di pubblicazione

Tramite i telefonini, inviate al Mentimeter le minacce che vi vengono in mente

An illustration within a circular frame showing a funnel at the top with various colorful icons (like a globe, smartphone, target, and game controller) falling into it. Below the funnel is a laptop computer with a blue screen, representing the filtered content being processed or displayed.

- Il filtraggio è un processo che implica il rilevamento e il blocco / rimozione di contenuti ritenuti inappropriati.
- I filtri possono essere impostati dai moderatori di una piattaforma o di un account di social media e possono essere percepiti come un insieme di regole a cui il contenuto si deve adeguare
- A tal fine, spesso sviluppano delle "blacklist" (un resoconto di parole, termini, forme di contenuto o account e siti di origine) che vengono bloccati

Il filtraggio è un processo che implica il rilevamento e il blocco / rimozione di contenuti ritenuti inappropriati. I filtri possono essere impostati dai moderatori di una piattaforma o di un account di social media e possono essere percepiti come un insieme di regole a cui il contenuto si deve adeguare per essere pubblicato o rimanere visibile tramite una piattaforma (Council d'Europa: 2017). A tal fine, spesso si sviluppano "black-list" (un resoconto di parole, termini, forme di contenuto o account e siti di origine) a cui viene impedito di pubblicare, condividere o commentare sulla piattaforma o sull'account.

- I moderatori dei contenuti compilano un archivio di parole o termini che non sono consentiti per la loro piattaforma, nonché un elenco di siti Web, account di posta elettronica o utenti che devono essere bloccati dalla pubblicazione o dalla condivisione sulla loro piattaforma
- Questo è un processo che può essere supportato da strumenti e plugin che possono svolgere questo lavoro automaticamente
- Tali strumenti supportano i moderatori nell'intraprendere da esperti questo compito e aggiornano regolarmente la loro 'black-list' / lista di blocco aggiungendo o rimuovendo termini, account, fonti di contenuto



In altre parole, i moderatori di contenuti possono compilare un archivio di parole o termini che non sono consentiti per la loro piattaforma, nonché un elenco di siti Web, account di posta elettronica o utenti che devono essere bloccati dalla pubblicazione o condivisione sulla loro piattaforma. Questo è un processo che può essere supportato da strumenti e plug-in che possono svolgere questo lavoro automaticamente, supportando i moderatori nell'intraprendere abilmente questo compito e mantenere i loro standard nelle loro piattaforme o account. Permette inoltre di aggiornare regolarmente la loro 'blacklist/ lista di blocco aggiungendo o rimuovendo termini, account, fonti di contenuto ecc.



Al giorno d'oggi ci affidiamo molto più spesso ai media di Internet per ricevere le notizie, rispetto alle fonti di notizie tradizionali.

I forum online ci consentono di condividere conoscenze, commenti, messaggi o di discutere un particolare argomento con altre persone con il nostro stesso interesse.

I social network o siti Web, come Facebook, Instagram, Twitter, YouTube, Tumblr, LinkedIn, Snapchat, Quora, Reddit, Pinterest, sono ampiamente utilizzati da persone in tutto il mondo e ci consentono di trovare e condividere lì sopra qualsiasi notizia.

I podcast consentono a chiunque di condividere le proprie conoscenze e di comunicare con il mondo attraverso una serie di audio incentrati su un particolare argomento o tema.

Strumenti di moderazione dei contenuti

La crescente necessità di moderazione dei contenuti condivisi attraverso le piattaforme online ha portato allo sviluppo di un'ampia gamma di strumenti e piattaforme per supportare i moderatori-amministratori nel processo.

Questi strumenti vanno dai processi di filtraggio guidati dall'intelligenza artificiale a semplici plug-in per aiutarli a mantenere un certo controllo e supervisione.

Infine, una piattaforma può scegliere di non disporre di alcuna forma di moderazione o monitoraggio. Ciò è sconsigliato in quanto la mancanza di moderazione, in pratica, significa che il proprietario (i) gestore (i) di una piattaforma non hanno alcun controllo sul contenuto che quindi potrebbe diventare una miriade di materiale e di opinioni inappropriati, minacciosi o illegali.

Strumenti di moderazione dei contenuti

Strumenti di moderazione dei commenti per i blogger

Disqus

Un plug-in che consente agli amministratori di rivedere e moderare i commenti, nonché di creare e gestire elenchi di ban (utenti vietati), filtri di parole, controlli antispam.

Commenti di Facebook

Un plug-in che consente ai propri utenti di collegare i propri commenti ai propri profili Facebook e consente ai moderatori di moderare i commenti tramite l'app Facebook o tramite un browser e applicare azioni collettive ai commenti.

Disqus

Disqus è uno strumento che può essere facilmente installato in un blog tramite un codice drop-in o come plug-in e consente agli amministratori di rivedere e moderare i commenti sugli articoli tramite un'unica dashboard. La dashboard fornisce loro anche la possibilità di creare, tra le altre cose, elenchi di utenti vietati, filtri di parole e controlli antispam.

Commenti di Facebook

I blog possono utilizzare il plug-in Commenti di Facebook che consente ai loro utenti di collegare i loro commenti ai loro profili Facebook senza dover registrarsi e accedere a detto blog.

Questo strumento consente ai moderatori di organizzare i commenti in base al momento in cui vengono pubblicati o in base al loro impegno. Ancora più importante, viene data loro la possibilità di moderare i commenti tramite l'app Facebook o tramite un browser e applicare azioni collettive ai commenti. Possono anche decidere se lasciare pubblici o nascondere i commenti contrassegnati da Facebook o da altri utenti (ad esempio perché offensivi o blasfemi). Infine, se la moderazione di un sito / blog viene intrapresa da un team, i moderatori possono anche assegnare commenti specifici a diversi membri del team.

Strumenti di moderazione dei contenuti

Strumenti di moderazione dei commenti per i blogger

IntenseDebate

Consente ai moderatori di rivedere e rispondere ai commenti tramite e-mail, nonché di suddividere il compito di moderazione all'interno di un team di amministratori. Si possono anche applicare filtri, cercare e / o eliminare automaticamente i commenti per parola chiave, indirizzo IP o indirizzo e-mail e, se necessario, bannare gli utenti.

Livefyre

Livefyre consente ai moderatori di rivedere i commenti prima che vengano pubblicati, consente agli utenti di modificarli come creare elenchi di utenti esclusi e regole per la gestione automatica dei commenti

IntenseDebate

IntenseDebate è uno strumento che può essere sincronizzato in una varietà di piattaforme di siti Web come Wordpress, Tumplr o Blogger. Consente ai moderatori di rivedere e rispondere ai commenti tramite e-mail, nonché di suddividere il compito di moderazione all'interno di un team di amministratori. Ancora più importante, le sue funzionalità forniscono la capacità ai moderatori di filtrare, cercare e / o eliminare automaticamente i commenti per parola chiave, indirizzo IP o indirizzo e-mail e, se necessario, bannare gli utenti.

Livefyre

Livefyre offre una varietà di impostazioni personalizzabili che aiutano a condurre automaticamente il processo di moderazione. I moderatori hanno la possibilità di rivedere i commenti prima che vengano pubblicati, consentire agli utenti di modificarli e hanno la possibilità di creare elenchi di ban che impediscono a utenti specifici di commentare. Inoltre, possono impostare regole per i commenti, in modo che se, ad esempio, un numero di utenti contrassegna un contenuto, questo viene automaticamente eliminato. Lo stesso vale per il filtro volgarità della rete.

Strumenti di moderazione dei contenuti

Strumenti per affrontare gli abusi sui social media - Twitter

Silenziare

Quando si riceve un abuso online tramite Twitter, è consigliabile che i moderatori disattivino l'audio invece di bloccare gli account in questione. Questo approccio mitiga l'impatto dell'abuso poiché l'account di destinazione non riceve notifiche dall'account disattivato.

Bloccare

Un'azione di ultima istanza contro gli account che inviano spam in modo persistente o che inviano contenuti offensivi.

Strumenti di moderazione dei contenuti

Strumenti per affrontare gli abusi sui social media - Twitter

Segnalare

I moderatori generalmente segnalano a Twitter tweet o account che diffondono minacce potenzialmente credibili e imminenti o contengono immagini violente.

Strumenti di moderazione dei contenuti

Strumenti per affrontare gli abusi sui social media - Facebook

Eliminare un commento

Facebook offre agli amministratori della pagina la possibilità di eliminare un commento che potrebbero percepire come offensivo, minaccioso o dispregiativo.

Nascondere un commento

I moderatori possono scegliere di nascondere semplicemente un commento offensivo da un post. Tuttavia, questo è considerato meno efficace rispetto all'eliminazione, perché ciò significa che il commento in questione rimane visibile all'utente e ai suoi amici.

Strumenti di moderazione dei contenuti

Strumenti per affrontare gli abusi sui social media - Facebook

Escludere un utente dalla pagina

Se un utente pubblica ripetutamente commenti che violano gli standard di una pagina, minando i valori di una sana discussione, si consiglia di bandirlo.

Disattivare / disattivare i commenti

Questa è una funzionalità disponibile solo sui post video ed è un approccio che i moderatori scelgono quando ritengono di non avere la capacità di monitorare con successo il flusso di commenti su un video o un live streaming.

Strumenti di moderazione dei contenuti

Strumenti per affrontare gli abusi sui social media - Facebook

Blocca parole

I moderatori possono sfruttare la forza del filtro volgarità della loro pagina vietando parole specifiche, assicurando così che tutti i commenti che li includono non vengano pubblicati nella loro pagina.

Segnalazione

Se un moderatore ritiene che un commento / post violi gli standard di Facebook, può segnalare l'utente o la pagina che lo ha creato.

Strumenti di moderazione dei contenuti

Strumenti / piattaforme di moderazione automatica dei contenuti

Akismet

Akismet è un plugin che si concentra principalmente sul blocco dello spam. Può anche essere utilizzato come strumento di scansione per monitorare post e pagine e controllare i commenti per garantire che qualsiasi contenuto dannoso non appaia su un sito. Per maggiori informazioni:
<https://akismet.com/>

Utopia AI Moderator

Utopia AI Moderator è uno strumento completamente automatizzato progettato per monitorare e sradicare contenuti offensivi, fraudolenti e spam. Consente agli amministratori di definire la loro politica di moderazione e può gestire i contenuti in più lingue. Per maggiori informazioni:
<https://utopiaanalytics.com/utopia-ai-moderator/>

Akismet

Akismet è un plug-in che funziona meglio nelle piattaforme basate su WordPress e si concentra principalmente sul blocco dello spam. Può anche essere utilizzato come strumento di scansione per monitorare post e pagine e controllare i commenti per garantire che qualsiasi contenuto dannoso non appaia su un sito. È importante notare che questo plugin è sia facile da usare che gratuito.

Per maggiori informazioni: <https://akismet.com/>

Utopia AI Moderator

Utopia AI Moderator è uno strumento di moderazione completamente automatizzato specificamente progettato per monitorare e sradicare contenuti offensivi, fraudolenti e spam. Lo strumento modera il 100% dei contenuti in arrivo e rimane aggiornato imparando mentre funziona esaminando le decisioni di pubblicazione che i moderatori umani hanno preso in passato. Consente agli amministratori di definire la loro politica di moderazione, che segue automaticamente e può gestire i contenuti in più lingue, elaborare collegamenti e indirizzi. Utopia AI Moderator è anche in grado di comprendere il contesto di un contenuto valutato, indipendentemente dal gergo e dall'ortografia, mentre offre la moderazione automatica di video e immagini.

Per maggiori informazioni: <https://utopiaanalytics.com/utopia-ai-moderator/>

Strumenti di moderazione dei contenuti

Strumenti / piattaforme di moderazione automatica dei contenuti

Implio

Uno strumento che supporta contenuti automatizzati e manuali nonché la moderazione dei commenti. Offre una moderazione manuale tramite un'interfaccia per il controllo e consente inoltre ai moderatori di impostare una vasta gamma di regole di automazione per semplificare l'intero processo.

Per maggiori info: <https://besedo.com/implio-features/>

PicPurify

PicPurify è un'API di moderazione delle immagini in tempo reale progettata per rilevare e filtrare automaticamente le immagini con contenuti indesiderati. È in grado di identificare elementi dannosi nelle immagini e i moderatori possono impostare i parametri desiderati. Per maggiori informazioni:

<https://www.picpurify.com/>

Implio

Implio è uno strumento che può essere utilizzato supportando contenuti manuali e automatizzati, nonché la moderazione dei commenti. Offre un'interfaccia di moderazione manuale per controllare il processo e consente agli utenti di impostare i propri filtri personalizzati per individuare e rimuovere facilmente i contenuti indesiderati con schemi prevedibili. Inoltre, consente loro di impostare una vasta gamma di regole di automazione per semplificare l'intero processo.

Per maggiori info: <https://besedo.com/implio-features/>

PicPurify

PicPurify è un'API di moderazione delle immagini in tempo reale progettata per rilevare e filtrare automaticamente le immagini con contenuti indesiderati. È in grado di identificare elementi dannosi nelle immagini come nudità, droghe, odio e garantire che non appaiano su una piattaforma. I moderatori sono in grado di creare un approccio su misura che si adatti alle loro esigenze e prospettive e gli strumenti funzionano 24 ore su 24 per garantire che i loro requisiti siano soddisfatti.

Per maggiori informazioni: <https://www.picpurify.com/>

Strumenti di moderazione dei contenuti

Strumenti / piattaforme di moderazione automatica dei contenuti

Automated Intelligent Moderation di WebPurify

Uno strumento che offre una protezione 24 ore su 24 dai rischi associati alle immagini generate dagli utenti rilevando e rimuovendo nudità e altri contenuti inappropriati in tempo reale. Per maggiori informazioni:

<https://www.webpurify.com/photo-moderation/automated/>

Moderatore dei contenuti di Azure

Un servizio cognitivo che monitora il contenuto in più forme come testo, immagine e contenuto video, controllando e applicando etichette (flag) appropriate per materiale che potrebbe essere offensivo, rischioso o altrimenti indesiderabile.

Per maggiori informazioni: <https://docs.microsoft.com/en-in/azure/cognitive-services/content-moderator/overview>

Automated Intelligent Moderation di WebPurify

Il servizio Automated Intelligent Moderation (AIM) di WebPurify offre una protezione 24 ore su 24 dai rischi associati alle immagini generate dagli utenti rilevando e rimuovendo ad esempio nudità e altri contenuti inappropriati in tempo reale. Gli strumenti possono rilevare immagini contenenti nudità, alcol, droghe, gesti offensivi e simboli e testi di odio per garantire che il contenuto in una piattaforma sia "al sicuro" da tali contenuti offensivi o fraudolenti. Inoltre, offre modelli di moderazione personalizzati che possono adattarsi e soddisfare al meglio le esigenze di ciascuna piattaforma.

Per maggiori informazioni: <https://www.webpurify.com/photo-moderation/automated/>

Moderatore dei contenuti di Azure

Azure Content Moderator è un servizio cognitivo che monitora il contenuto in più forme come testo, immagine e contenuto video, controllando e applicando etichette (contrassegni) appropriate per il materiale che potrebbe essere offensivo, rischioso o altrimenti indesiderabile. Il servizio Content Moderator include anche lo strumento di revisione basato sul Web, che ospita le revisioni dei contenuti affinché i moderatori umani possano elaborarle. Combinando il lavoro del servizio con i team di revisione umani, i moderatori della piattaforma possono trovare il giusto equilibrio tra efficienza e precisione. Lo strumento Revisione fornisce anche un front-end intuitivo per diverse risorse di Content Moderator.

Per maggiori informazioni: <https://docs.microsoft.com/en-in/azure/cognitive-services/content-moderator/overview>



The graphic features a dark blue background with the text "Social Media Moderation" in white and orange at the top left. It includes various social media icons such as Facebook, Twitter, Instagram, and YouTube, along with communication symbols like speech bubbles, a heart, and an envelope. A hand is shown pointing at the Facebook icon. The entire graphic is partially enclosed by a white curved line on the right side.

Attività Nr 3
Presentazione di
gruppo su esempi
di moderazione
sui social media

Proprietà intellettuale.

Fonte video: <https://www.youtube.com/watch?v=UqZJPuyK9VY>

La necessità della moderazione dei contenuti nei social media

- Le piattaforme di social media sono diventate le più importanti fonti di contenuti e informazioni.
- Sebbene ciò abbia determinato una rivoluzione nel modo in cui le notizie e le informazioni vengono consumate, condivise e deliberate, esse hanno anche permesso la diffusione di disinformazione, odio, cyberbullismo e contenuti dannosi.
- Di conseguenza, c'è stata una crescente domanda da parte di piattaforme di social media per agire contro tali forme di contenuto, portandole a prendere provvedimenti per il monitoraggio e la moderazione.

Le piattaforme di social media sono diventate le più importanti fonti di contenuti e informazioni. Una grande percentuale di utenti Internet si affida ai social media per accedere a informazioni, notizie e punti di vista, nonché per interagire con i propri colleghi. Mentre questo ha portato a una rivoluzione nel modo in cui le notizie e le informazioni vengono consumate, condivise e deliberate, le piattaforme social media hanno anche permesso la diffusione di disinformazione, odio, cyberbullismo e contenuti dannosi.

Di conseguenza, c'è stata una crescente domanda di piattaforme di social media per agire contro tali forme di contenuto, portandole a prendere provvedimenti per il monitoraggio e la moderazione.

Risposta delle piattaforme di social media

- Le piattaforme di social media hanno sviluppato standard comunitari che delineano cosa è accettabile e cosa non lo è.
- Stanno impiegando strategie elaborate di moderazione dei contenuti che coinvolgono sia moderatori umani che sistemi automatizzati per garantire che i loro utenti siano protetti, nella misura in cui è possibile, da contenuti irriverenti e abusi.
- Chiedono agli utenti di segnalare contenuti che potrebbero offendere o che violano le regole della piattaforma.

Le piattaforme di social media sono diventate le più importanti fonti di contenuti e informazioni. Una grande percentuale di utenti di Internet si affida ai social media per accedere a informazioni, notizie e punti di vista, nonché per interagire con i propri colleghi. Se da una parte questo ha portato a una rivoluzione nel modo in cui le notizie e le informazioni vengono consumate, condivise e deliberate, le piattaforme di social media hanno anche permesso la diffusione di disinformazione, odio, cyberbullismo e contenuti dannosi.

Di conseguenza, c'è stata una crescente domanda di piattaforme di social media per agire contro tali forme di contenuto, portandole a prendere provvedimenti per il monitoraggio e la moderazione.



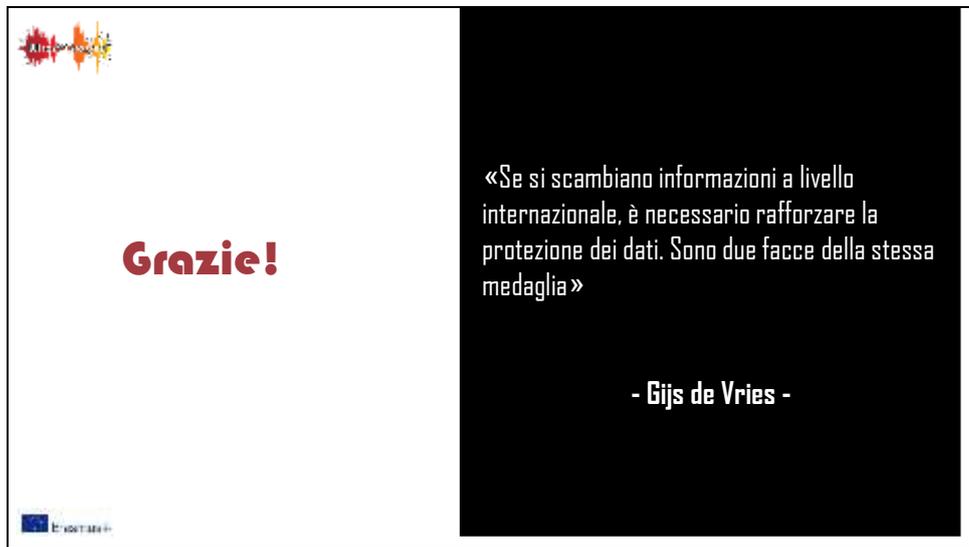
Attività Nr3: presentazione di gruppo
su esempi di moderazione dei social

Attività Nr 4: Discussione sulle minacce su Internet

paloalto | UNIT



Quando si tratta di moderazione dei contenuti nei social media, si comincia a discutere di libertà di parola e della paura della censura o del controllo delle opinioni condivise. Guarda questo video come introduzione ad una discussione che può avvenire anche all'interno del gruppo





Grazie!



«Se si scambiano informazioni a livello internazionale, è necessario rafforzare la protezione dei dati. Sono due facce della stessa medaglia»

- Gijs de Vries -

Grazie!



The logo for 'Wires-Crossed' features the text 'Wires-Crossed' in a bold, white, sans-serif font. The text is centered and overlaid on a background of numerous thin, black lines that radiate outwards from the text, creating a starburst or network effect. The lines are colored in shades of red and orange, with the red being more prominent on the left and the orange on the right.



A row of seven partner logos is displayed below the main title. From left to right: 'dante' (purple square with white text), 'AIK' (red square with white text), 'Speha Fresia' (red square with white text and a stylized 'S' logo), 'JUGEND- & KULTURPROJEKT E.V.' (yellow circle with a black 'i' logo), 'The Rural Hub' (blue and green logo with a stylized 'R' and 'H'), 'CARDET' (blue logo with a stylized 'C' and 'D' logo), and 'ACUMEN TRAINING' (red logo with a white 'A' and 'M' logo).



The logo for the Erasmus+ Programme of the European Union, featuring the European Union flag (a blue rectangle with twelve yellow stars arranged in a circle) and the text 'Co-funded by the Erasmus+ Programme of the European Union'.

Co-funded by the
Erasmus+ Programme
of the European Union

The European Commission's support for the production of this publication does not constitute an endorsement of the contents, which reflect the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein. 2019-1-DE02-KA204-006115