|  |  |
| --- | --- |
| Comprensione del comportamento scorretto nella discussione di COVID19 sui social network e sui media online | |
|  | |
| Studente/i | Relatore |
| Nogara Gianluca | Giordano Cremonese Silvia |
| Correlatore |
| Luceri Luca |
| Committente |
| - |
| Corso di laurea | Modulo |
| Ingegneria Informatica TP | C10344 Progetto di diploma |
| Anno |  |
| 2020/2021 |  |

|  |  |
| --- | --- |
| Data |  |
| - |  |

**Indice**

[Abstract 7](#_Toc71295617)

[Progetto assegnato 9](#_Toc71295618)

[1 Analisi 11](#_Toc71295619)

[1.1 Contesto 11](#_Toc71295620)

[1.2 Fruitori 11](#_Toc71295621)

[1.3 Analisi delle tecnologie 11](#_Toc71295622)

[1.4 Requisiti 12](#_Toc71295623)

[1.4.1 Contenuti 12](#_Toc71295624)

[1.4.2 Funzionalità 12](#_Toc71295625)

[1.4.3 Rispetto della privacy 13](#_Toc71295626)

[1.5 Obiettivi 13](#_Toc71295627)

[2 Studio delle Soluzioni 14](#_Toc71295628)

[2.1 La nostra scelta 14](#_Toc71295629)

[2.2 Moodle plugin 14](#_Toc71295630)

[2.3 L’alternativa 14](#_Toc71295631)

[2.4 Learning Locker 15](#_Toc71295632)

[2.5 Ambienti di sviluppo e linguaggi 16](#_Toc71295633)

[2.5.1 Anaconda e il linguaggio Python 17](#_Toc71295634)

[2.5.2 MariaDB 17](#_Toc71295635)

[2.5.3 Bash 17](#_Toc71295636)

[2.5.4 Postman 17](#_Toc71295637)

[2.6 Strumenti utilizzati 18](#_Toc71295638)

[2.6.1 VMware 18](#_Toc71295639)

[2.6.2 PuTTY 18](#_Toc71295640)

[3 Implementaizone 19](#_Toc71295641)

[4 Operazioni sul LRS ed API 22](#_Toc71295646)

[4.1 customQuery.py 22](#_Toc71295647)

[4.2 LLSender.py 24](#_Toc71295648)

[4.3 MSTeams 26](#_Toc71295649)

[5 Piani di lavoro 28](#_Toc71295650)

[5.1 Piattaforma SCM 28](#_Toc71295651)

[5.2 Repository 28](#_Toc71295652)

[6 Conclusioni 29](#_Toc71295653)

[6.1 Problemi riscontrati 29](#_Toc71295654)

[6.2 Implementazioni future 30](#_Toc71295655)

[7 Fonti 31](#_Toc71295656)

**Indice delle figure**

[Figura 1: Esempio di uno Statement xAPI 12](#_Toc71295657)

[Figura 2: Esempio di dati anonimizzati nel database di Moodle 13](file:///\\win1.dti.supsi.ch\ProgettiDiStudio2021\ProgettoSem\Info\Moro%20-%20Nogara\Relazione.docx#_Toc71295658)

[Figura 3: Pagina di login di LL 15](#_Toc71295659)

[Figura 4: Pagina principale di LL, sezione Data > Dashboards 16](#_Toc71295660)

[Figura 5: Configurazione PuTTY 18](#_Toc71295661)

[Figura 6: Stato dei servizi di LL dopo l'installazione via script 20](#_Toc71295662)

[Figura 7: Esempio di una post tramite Postman 21](#_Toc71295663)

[Figura 8: Schema della raccolta e normalizzazione dei dati 22](#_Toc71295664)

[Figura 9: Schema del fetiching e inserimento in Learning Locker 24](#_Toc71295665)

[Figura 10: Risultato dell'inserimento visto all'interno di Learning Locker 26](#_Toc71295666)

[Figura 11: Esempio di dati prima del parsing 27](#_Toc71295667)

[Figura 12: Esempio di dati dopo il parsing 27](#_Toc71295668)

[Figura 13: Organizzazione del lavoro 28](#_Toc71295669)

[Figura 14: Sprint Board iterazione-05 29](#_Toc71295670)

Abstract

I social network e i media online (OSNEM) sono tecnologie basate su computer che consentono agli utenti di creare contenuti e intrattenere relazioni sociali.

Le OSNEM sono rapidamente cresciute dall’essere semplicemente un canale di aggregazione all’essere un fenomeno globale che ha suscitato un cambio di paradigma della nostra società, trasformando il modo in cui le persone accedono alle notizie, condividono opinioni, fanno affari e fanno politica.

In uno scenario del genere, l'accuratezza, la veridicità e l'autenticità del contenuto condiviso sono ingredienti necessari per mantenere una sana discussione online.

Tuttavia, negli ultimi tempi, le OSNEM hanno dovuto affrontare una notevole crescita di account e attività dannose, che minano intenzionalmente l'integrità delle conversazioni online condividendo, nelle piattaforme OSNEM, informazioni false e provocatorie per influenzare l'opinione pubblica e creare conflitti su questioni sociali o politiche.

Ciò è stato, e continua ad essere, drammaticamente vero per la discussione sul COVID19, che è stata minata da campagne di manipolazione e disinformazione a livello globale. È quindi necessario studiare i comportamenti peculiari delle entità dannose nel dibattito COVID19 per fornire un cambiamento sociale verso una migliore comprensione (alfabetizzazione) delle reti sociali.

Questo include il rilevamento di malintenzionati, l'identificazione di informazioni di scarsa credibilità e notizie false, la consapevolezza degli utenti e il controllo dei comportamenti scorretti.

Social networks and online media (OSNEMs) are computer-based technologies that allow users to create content and engage in social relationships.

OSNEMs have rapidly grown from being simply a channel of aggregation to being a global phenomenon that has sparked a paradigm shift in our society, transforming the way people access news, share opinions, do business, and make policy.

In such a scenario, accuracy, truthfulness and authenticity of shared content are necessary ingredients to maintain a healthy online discussion.

However, in recent times, OSNEMs have faced a significant growth of malicious accounts and activities, which intentionally undermine the integrity of online conversations by sharing, in OSNEM platforms, false and provocative information to influence public opinion and create conflict on social or political issues.

This has been, and continues to be, dramatically true of the COVID19 discussion, which has been undermined by global campaigns of manipulation and misinformation. Therefore, it is necessary to study the peculiar behaviors of malicious entities in the COVID19 debate to provide social change towards a better understanding (literacy) of social networks.

This includes detection of malicious actors, identification of information of low credibility and fake news, user awareness, and control of misbehavior.

Progetto assegnato

Lo scopo di questo progetto è quello di analizzare entità malevoli presenti in un dataset pubblico su COVID19. L’ OSNEM di riferimento è Twitter e l’attività di analisi e classificazione delle entità malevoli (come bot e trolls) viene fatta tramite diversi algoritmi esistenti.

Il progetto in questione consente quindi di analizzare il comportamento all'interno della discussione e l’impatto che hanno tali entità malevole nella discussione globale.

**COMPITI**

•

•

•

**OBIETTIVI**

L'obiettivo principale di questa tesi è quello di arrivare a distinguere tra entità/azioni corrette e quelle malevoli nel dibattito COVID19.

Questo include il rilevamento di malintenzionati (ad esempio bot e troll), l'identificazione di informazioni di scarsa credibilità e notizie false, la consapevolezza degli utenti e il controllo dei comportamenti scorretti.

**TECNOLOGIE**

• Python

• Complex networks

# **1 Analisi**

## 1.1 Contesto

Il progetto si inserisce nel contesto del Dipartimento Tecnologie Innovative della SUPSI e in particolare nel corso di Ingegneria Informatica. Si tratta di un progetto interno alla SUPSI, anche se di grande importanza sociale dal momento che le informazioni raccolte possono essere utilizzate per limitare la propagazione di entità malevoli sui social network.

## 1.2 Fruitori

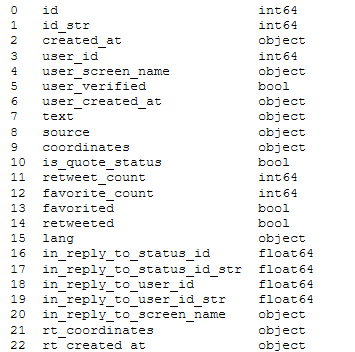
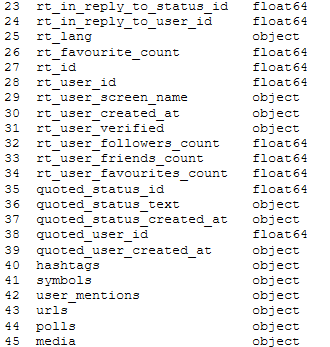
## 1.3 Analisi delle tecnologie

Il primo compito assegnato riguarda l’analisi di un sample dei dati (di soli 50.000 campioni) in modo da poter familiarizzare con la struttura del dataset e operare in correttamente sui dati.

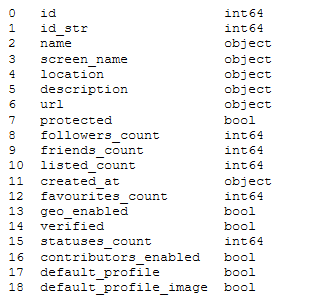
I dataset in questione riguardano i profili utente e i singoli tweet, la struttura dei due dataset differisce per alcuni campi, in particolare quello dedicato agli utenti è molto ricco di campi (ben 46), mentre quello dedicato agli utenti ne ha poco meno della metà (18).

I dati da analizzare risultano eterogenei, sono presenti booleani, valori interi, oggetti e valori in virgola mobile, è quindi importante il processamento di tali per un corretto studio.

Tweets



E users



## 1.4 Requisiti

### 1.4.1 Contenuti

Il principale requisito a livello contenutisco è quello di avere dei feedback visivi dell’analisi effettuata sui dati, in modo da effettuare operazioni di ulteriore processamento realizzando delle …

### 1.4.2 Funzionalità

Per quando riguarda i requisiti legati all’usabilità del sistema sviluppato possiamo identificare come utente finale sia un’organizzazione che un singolo individuo. Infatti, essendo il progetto open source e completamente personalizzabile, è possibile implementarne funzionalità ad-hoc.

### 1.4.3 Rispetto della privacy

Dal momento che i dati utilizzati dal LRS sono informazioni di studenti e docenti si è deciso di anonimizzare i dati sensibili, come il nome e il cognome di docenti e studenti. Questo ha portato ad avere delle informazioni aggregate utilizzabili ma totalmente privacy-friendly, come riportato nella Figura 2.

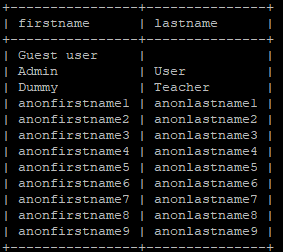


Figura 2: Esempio di dati anonimizzati nel database di Moodle

## 1.5 Obiettivi

Gli obiettivi di questo progetto sono i seguenti:

* Installare e configurare un LRS interagibile su una macchina Unix
* Accedere ai dati di una piattaforma Moodle di prova con dei dati anonimizzati
* Accedere ai dati tool di videoconferenza (p.e. MS Teams)
* Implementare delle API per estrarre i dati e importarli nel LRS
* Visualizzare i dati tramite un tool di visualizzazione commerciale (p.e. Tableau)

# 2 Analisi delle Soluzioni

## 2.5 Ambienti di sviluppo e linguaggi

Dal momento che tra gli obiettivi figuravano lo sviluppo di API, vi era la necessità di utilizzare un linguaggio flessibile che potesse gestire diversi tipi di dati in modo semplice e veloce, interagendo inoltre con endpoint e database.

La scelta di un linguaggio che rispettasse i requisiti sopra citati ha orientato l’attenzione verso Python.

### 2.5.1 Anaconda e il linguaggio Python

Anaconda ha svolto un compito molto importante per quanto riguarda la programmazione in Python, infatti, dal momento che per lo sviluppo delle API vi è l’esigenza di poter effettuare operazioni di mapping, andando a sostituire dei placeholder da un file di sorgente con dei dati custom.

Ad esempio, ci sono API che richiedono di effettuare delle query personalizzabili dagli utenti verso un database, richiedendo quindi di inserire dei dati in un file

Uno strumento essenziale è stato il Notebook, che ha permesso di gestire in modo ottimale i programmi, garantendo una buona gestione del codice.

### 2.5.2 MariaDB

MariaDB è un DBMS, non è altro che una fork di MySQL ed è stato largamente utilizzato per poter accedere ai dati anonimizzati caricati sulla piattaforma Moodle di prova.

L’esigenza dell’utilizzo di tale database è data dal momento che è stato utilizzato un database di Moodle di prova, basato su MariaDB, che conteneva dati anonimizzati di prova che andavano importati sul LRS implementato nelle fasi precedenti.  
Il database è stato esplorato per poterne capire la struttura e tipi di dati contenuti, inoltre abbiamo potuto effettuare delle query per poter testare l’efficacia dei programmi realizzati. Per l’accesso a MariaDB stato effettuato uno script in modo da poter entrare senza dover esplicitamente scrivere differenti linee di codice ogni volta.

## 2.6 Strumenti utilizzati

Per lavorare al LRS è stato necessario organizzare il lavoro in modo da poter accedere a una singola macchina in remoto accessibile da entrambi gli studenti.

Prima di accedere alla macchina in remoto, per evitare qualsiasi rischio è stato svolto un lavoro di implementazione in locale usando delle macchine virtuali.

### 2.6.1 VMware

Come anticipato si è adottata una strategia di contenimento dei danni, sono state sfruttate al meglio le caratteristiche delle macchine virtuali: garantiscono un ambiente completamente isolato e ripristinabile, quindi si è testata l’installazione del LRS prima in un ambiente virtuale, scelta che si è rivelata molto importante.

### 2.6.2 PuTTY

Dovendo operare su una macchina in remoto si è usato un client con la funzione di emulatore di terminale per accedere tramite della macchina.

# 3 Implementazione

### 3.1.1 Problemi di prestazioni

### 3.1.2 Problemi di archiviazione

# 4 Operazioni sui dati

# 5 Piani di lavoro

## 5.1 Piattaforma SCM

Nei requisiti iniziali del relatore, era stato richiesto l’utilizzo di SCRUM (SCM) per lo sviluppo del progetto. Abbiamo organizzato la pianificazione del lavoro in diversi Sprint, in modo tale da tenere traccia delle attività svolte e mantenere l’organizzazione nelle operazioni

Sono stati effettuati sette Sprint, di seguito riportiamo le Issues con i relativi iterazioni.



Figura 13: Organizzazione del lavoro

## 5.2 Repository

Il repository è stato uno strumento cruciale per quanto riguarda la sincronizzazione del gruppo e il mantenimento del codice aggiornato e ordinato.

Il lavoro effettuato tramite Git è iniziato il 06/04/2021, con l’iterazione-05, questo perché le prime cinque iterazioni riguardavano la ricerca di informazioni, l’installazione di LL e la familiarizzazione con il software in questione.

Le attività hanno avuto una cadenza settimanale e bisettimanale in funzione del lavoro svolto e delle Issues aperte non ancora svolte.

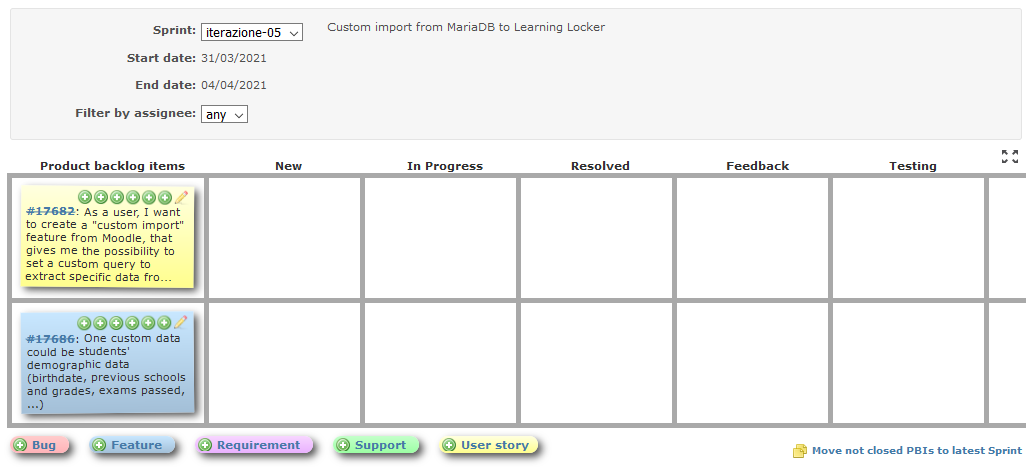


Figura 14: Sprint Board iterazione-05

# 6 Conclusioni

## 6.1 Problemi riscontrati

Purtroppo, durante lo sviluppo del progetto sono stati innumerevoli i problemi che sono stati riscontrati, tutti princiaplente sull’LRS:

* **Poco supporto** da parte dell’azienda che ha prodotto il software, nonostante la documentazione fosse buona, sono diverse le issues aperte e le domande che molti utenti si pongono.
* **I video dimostrativi sono ormai datati** e le tecnologie sono in continuo miglioramento, sarebbe opportuno aggiornare le informazioni e dare maggiore supporto agli sviluppatori.
* **Errori nell’installazione**, infatti durante la fase di installazione, anche se l’esecuzione dello script è stata corretta, il software contiene degli errori in fase di installazione che ha dovuto correggere il Docente.
* **API praticamente inesistenti**, infatti accedendo al sito di LearningPool le API esistono ma non possono essere utilizzate in alcun modo, oltre ai contatti per problemi possiamo cliccare su un bottone con scritto “Request a demonstration” che però non fa nulla.

## 6.2 Implementazioni future

Il progetto è destinato ad essere sviluppato, sarà parte infatti di un progetto di diploma.

Il tool attualmente è in grado di raccogliere i dati di log da iCorsi, inserire delle query personalizzate da iCorsi, e integrare i dati da Teams.

Sarà opportuno integrare dati da social network, come ad esempio Twitter, per ottenere e incrociare altri dati.

Si potrà implementare la possibilità di definire delle nuove metriche, partendo dai dati esistenti, per l’analisi dei dati.

Si potrebbero implementare altre API per estrarre i dati tramite tool di analisi visuale, come Tableau.

Sono innumerevoli inoltre le API che, in funzione delle esigenze, potrebbero essere implementate.

Oltre all’estensione come progetto di diploma è possibile estenderlo a chiunque ne abbia il bisogno, si potrebbe personalizzare per un determinato istituto o per un singolo docente che ha l’esigenza di processare dati di apprendimento.

# 7 Fonti

**Python:**

<https://botometer.osome.iu.edu/api>

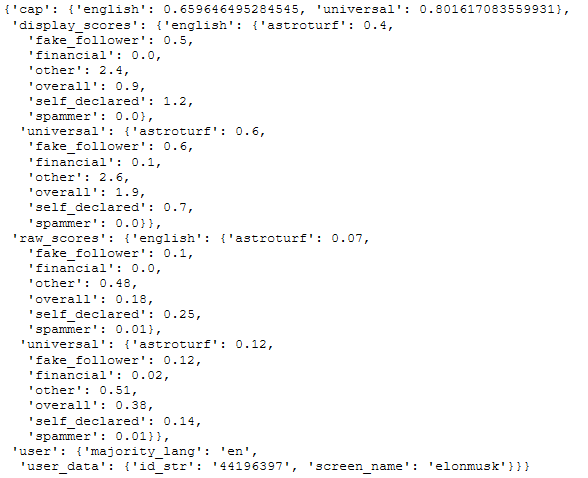
<https://pypi.org/project/botometer/#description>

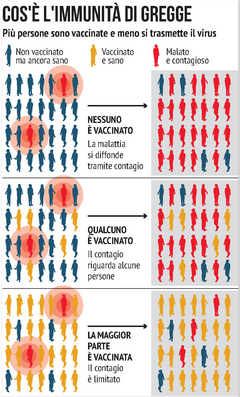
<https://cnets.indiana.edu/blog/2020/09/01/botometer-v4/>

<https://blog.quantinsti.com/detecting-bots-twitter-botometer/>

<https://rapidapi.com/developer/dashboard>

<https://networkx.org/documentation/stable/index.html>





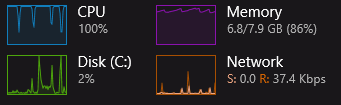
Problemi:

Memoria (parallelizzazione)

bottleneck (lettura in chunk)

Tipi di dati (parsing in date)

ParserError: Error tokenizing data. C error: out of memory



Confronto dei tempi

Da 1000+ secondi a 390 + secondi (users)

