Alma Mater Studiorum · Università di Bologna

SCUOLA DI SCIENZE
Corso di Laurea in Informatica per il Management

Rilevazione di eventi geo-localizzati sulla piattaforma Twitter: Una valutazione sperimentale

Relatore:
Chiar.mo Prof.
MARCO DI FELICE

Presentata da:
DAVIDE BACCA

Sessione III
Anno Accademico 2015/2016

Ai miei amici
"Nuèter a sän un quèl saul
#WeAreOne"
slogan Bologna Fc 1909

Introduzione

I Social Network si stanno diffondendo sempre di più nella società moderna. Essi rappresentano una grande fonte di informazioni istantanee e precise. Il Social Sensing, che é alla base di questo lavoro, si basa sull'idea che comunità o gruppi di persone possano fornire informazioni, su eventi o problemi che accadono nelle loro vicinanze, simili a quelle ottenibili da sensori. Fra i vari campi di studio ai quali si applica il Social Sensing, quello riguardo alla gestione delle emergenze tramite l'uso di informazioni derivanti dagli utenti dei Social Network, attraverso le quali si possono avere informazioni dettagliate e aggiornate su situazioni emergenti, di pericolo e non, per poter allertare le persone coinvolte e le autorità di competenza, é uno dei più interessanti. Ma l'utilità di queste informazioni non si limita all'allerta tempestiva, esse possono essere adoperate per coordinare le operazione di assistenza dopo l'accadere dell'evento o per prevedere situazioni collaterali dovute all'accadere di un fatto.
Sono molte le ricerche, presenti in letteratura, nell'ambito dell'utilizzo dei Social Media nel campo del rilevamento delle situazioni di emergenza. Nel Primo Capitolo si andranno ad esporre i principali lavori svolti a riguardo, essi sono stati presentati in base agli step comuni di analisi. Nel Secondo Capitolo si andrà ad esporre la piattaforma Twitter, utilizzata per la realizzazione del nostro sistema, descrivendone le origini e le API messe a disposizione dalla stessa per fruire dei contenuti in essa raccolti. Nel Terzo Capitolo si descrive, invece, il sistema da noi creato, esponendo la struttura dei moduli nei quali é suddiviso e le simulazioni fatte per testarne il funzionamento.

Il nostro progetto ha l'obbiettivo di dimostrare quanto possa essere semplice la creazione dello scheletro di una applicazione che, attraverso il controllo di Twitter, monitora l'accadere di eventi nella regione Emilia Romagna. Il nostro sistema, che prende spunto da lavori precedenti, rileva i tweet geo-localizzati e viene allertata inizialmente da una perturbazione del numero di messaggi composti dagli utenti per poi andare ad analizzare, più specificatamente, la distribuzione spaziale per appurare l'accadere di un evento.

Indice

Elenco delle figure

Elenco delle tabelle

Capitolo 1

Stato dell'arte

1.1 Stato dell'arte

I Social Media (SM) stanno assumendo una rilevanza sempre maggiore in molti ambiti della nostra vita quotidiana. In particolare, possono essere considerati una miniera di informazioni preziose e con tre caratteristiche rilevanti:

- Sono per la maggior parte di natura spontanea, e non guidata
- Sono provvisti di posizione spaziale (ad esempio, tramite tag)
- Sono prodotti in tempo reale.

La mole di dati prodotta è sempre maggiore. I SM sono un nuovo campo di applicazione delle analisi di tipo BigData e tali analisi possono essere riassunte con lo schema 4V (Varietà, Volume, Velocità, Valore) [21]. I SM si pongono, quindi, come un interessante strumento di studio degli eventi. In particolare, negli ultimi anni l'attenzione si è focalizzata sul rilevamento e la gestione di emergenze, intendendo, per emergenza, un evento improvviso che cambia il corso naturale degli eventi[21]. I tentativi di analisi e gestione dei dati estrapolati dai SM per la gestione di eventi non previsti (incendi, uragani, terremoti) sono molteplici. Sono generalmente basati sul presupposto

che un gruppo di persone possa fornire un set di informazioni paragonabili a quelle fornite da un singolo sensore e hanno come scopo il poter rilevare tempestivamente eventi di preoccupazione sociale [2]. La letteratura in merito è sempre più ampia.

I tentativi di analisi dei SM presentano una struttura comune:

- Acquisizione dei dati
- Filtraggio (pre-processamento e processamento) dei dati
- Classificazione dei dati
- Conclusioni

La scelta del SM su cui basare uno studio risulta importante. La piattaforma Twitter presenta una serie di vantaggi rispetto agli altri SM. Twitter è, infatti, un sistema di micro-blogging che permette la scrittura di testi (con tag o hashtag) limitati a 140 caratteri. L'acquisizione dei dati e la conseguente analisi sono rese dunque considerevolmente più snelle. Twitter può a tutti gli effetti essere considerato un sistema di Crowdsourcing: un gruppo di persone, organizzate o non organizzate, che forniscono informazione ad un'altra persona o ad un altro ente[10].

1.1.1 Acquisizione dei dati

Il processo di acquisizione viene effettuato mediante Twitter API. Diversi studi sono stati condotti utilizzando le Search API, tuttavia si possono utilizzare le Streaming API che consento l'accesso all'intero flusso di informazione, Twitter fornisce al massimo l'1% del traffico totale e taglia fuori l'eccesso.

Nel processo di acquisizione si possono inserire parole chiave e restringere il campo di ricerca richiedendo dati di geolocalizzazione in una distanza variabile a partire da una posizione specifica. Nel caso vengano estratti dati da social network diversi, si renderà necessario memorizzare i dati in un formato comune. Diversi sono i punti critici rilevati in fase di acquisizione.

In particolare, è stata evidenziata la necessità di una analisi rapida dei dati forniti dalle Streaming API poiché durante l'elaborazione il client non è in grado di accettare altri Tweet.

L'uso delle Streaming API non consente di estrapolare i Tweet prodotti prima della connessione, viene evidenziato come questo non rappresenti un problema, dal momento che quando utilizziamo API in streaming siamo interessati ad un lavoro su dati che emergono in tempo reale[7]. Alcuni studi possono non presentare la necessità di rilevare solo Tweet in tempo reale, ad esempio per lo studio delle condizioni climatiche, in cui l'estrazione dei Tweet veniva effettuata con cadenza giornaliera, si è potuto utilizzare Twitter4j Java Library, effettuando la ricerca, sia a partire da parole chiave specifiche, sia attraverso tag spaziali[19].

Nel lavoro di Avvenuti et al(2013)[2] l'evento oggetto di studio può essere rilevato in modi diversi. Viene proposto un rilevamento dell'evento sulla base di due caratteristiche, l'analisi temporale (lo scopo è rilevare un evento in tempo reale) e l'analisi spaziale. L'improvviso aumento di numero di Tweet 1.1, se superano una soglia critica data, diventa il primo fattore per il rilevamento di un evento di emergenza. L'analisi spaziale ha posto in evidenza il problema che, dei Tweet analizzati, solo l'1,5% era georeferenziato. Per ovviare a questo problema si è pensato di utilizzare la posizione dell'account di Twitter, cioè quella che l'utente può inserire alla registrazione, ma purtroppo questa potrebbe essere solo forviante in quanto non obbligatoria al momento della registrazione e non soggetta ad un controllo in tempo reale. Hanno utilizzato le Streaming API con le quale i nuovi Tweet che contengono le parole chiave possono essere raccolti. Streaming API permette, potenzialmente, di raccogliere tutti i Tweet che corrispondono ai criteri inseriti. Le parole chiave che utilizzano sono principalmente: terremoto e scossa.

Dong et al (2013)[7] studiano l'uso di Twitter per estrapolare informazioni utili dai sentimenti degli utenti in situazioni disastrose, in particolare durante il cataclisma dell'uragano Sandy. Per l'acquisizione dei dati utilizzano le Twitter Search API . I dati raccolti sono stati prodotti utilizzando il seguente

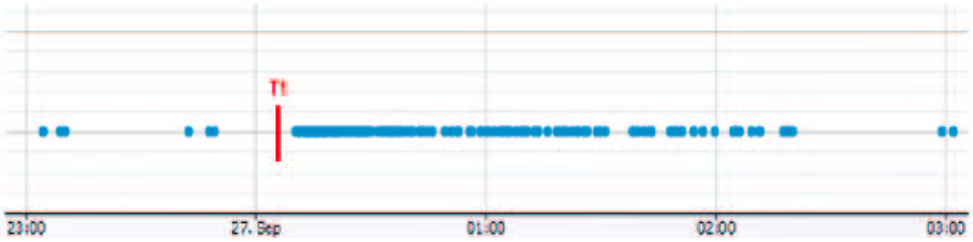

Figura 1.1: Un picco di Tweet registrato dopo un terremoto

set di parole chiave: "Evacuate, Prepare, Flood, Hurricane, Storm, Tornado, Sandy, Rain, Water, Food, Emergency, Battery, Safe, Leave". Salvati i Tweet in file text per evitare problemi nel processamento, hanno rimosso i caratteri speciali da ogni Tweet e hanno rimosso i Tweet doppi, causati dalle Search API che occasionalmente possono ridare il medesimo Tweet più volte. In [8] vengono sempre utilizzate le Search API per la raccolta dei dati, usando però come parole chiave gli hashtag : #kashmirfloods e #jammufloods.

In [10] i Tweet sono estratti giornalmente usando Twitter4j Java library. Durante la prima parte del progetto erano estratti solo una volta al giorno, subito prima di mezzanotte. Venivano eseguite due query. La prima era per trovare i Tweet etichettati in modo specifico, i tag utilizzati per la ricerca erano del tipo: #ptaweather, @PretoriaZA, #pretoria, #pta e #weather. La seconda si basava su latitudine e longitudine di Pretoria e cercava termini specifici, senza tag ma solitamente correlati a condizioni climatiche.

1.1.2 Processamento e classificazione dei dati

Una volta acquisiti i dati il passaggio successivo è il processamento. Esso consiste in una serie di tecniche volte a raffinare i dati ed è definito in base all'obbiettivo della ricerca; In[1] hanno identificato due tipi di "rumore" che vanno a interferire con la raccolta dei Tweet di interesse, i messaggi nei quali le parole chiave sono usate con significato differente da quello di interesse e i messaggi nei quali tali parole si riferiscono ad un evento passato. Il modulo Data Filtering (modulo di filtraggio dati) riduce questo "rumore"

pulendo i dati in due passaggi. Una fase di pre-filtraggio nella quale vengono eliminati i ReTweet, i messaggi di risposta, i messaggi provenienti da 345 account che appartengono a canali ufficiali e i messaggi che contengono una serie di parole inserite nella blacklist. Dopo questa fase si ha un filtraggio più specifico effettuato da WEKA che deduce la classe di appartenenza di un Tweet in base ad un modello. I dati "training", utilizzati per creare il classificatore sono divisi in due categorie: utili e non utili.

In[7] i dati riportati sono quelli dei Tweet scritti in prima persona, tali Tweet devono contenere parole chiave come "evacuate", "leave" o "go", indicatori del fatto che l'utente intende evacuare oppure "stay,prepare,power, battery, water, precaution", parole chiave che indicano l'intenzione di rimanere. Queste operazioni permettono, per ogni Tweet, la riduzione del cosiddetto "rumore di fondo". Risulta chiaro che le regole di rimozione vengono impostate in modo diverso a seconda dello scopo dello studio. I dati raccolti

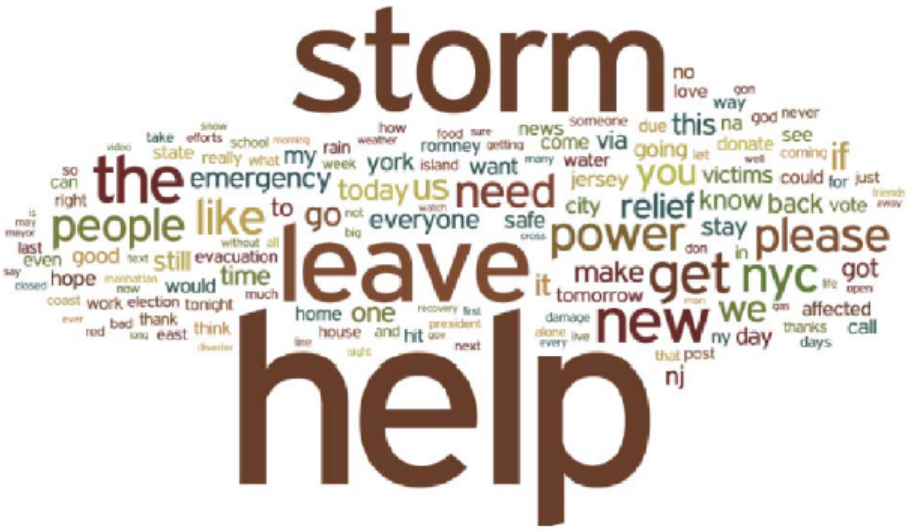

Figura 1.2: Word Cloud generato dal modello bag-of-words

dalle diverse aree colpite dall'uragano Sandy sono stati uniti in un unico file. Si sono andate a formare, quindi, delle "word cloud" (per esempio si riporta alla figura 1.2), usando il modello bag-of words, le quali includono più di 2000 parole nell'intero file per l'uragano Sandy. Lo scopo di questa "word

cloud" è di facilitare la classificazione dei dati. In questo lavoro il criterio di classificazione è la discriminazione tra coloro i quali intendono evacuare e coloro che invece restano nelle proprie case, in base ai loro Tweet. Per fare ciò usano l'algoritmo Latent Semantic Indexing (LSI) in gensim. Kaur e Kumar [3] nel set di dati sul kashmir hanno molti Tweet in Hindi tradotti in inglese usando le Google Translate API. I dati sono processati come di seguito: convertono i Tweets in minuscolo per mantenere il dataset uniforme; tutti gli URLs e User Names usati nei Tweets sono rimpiazzati con stringhe costanti; sono rimossi tutti gli hash tags (e.g. #kashmirfloods con kashmirfloods) così come gli spazi bianchi extra, i simboli speciali, la punteggiatura, i caratteri alpha-numerici, le parole comuni (common stop words) e i caratteri ripetuti in una parola. I dati sono poi divisi in tre categorie: negativi, positivi e neutri (come mostrato in figura 1.3). Come classificatore utilizzano Naive Bayes Classifier.

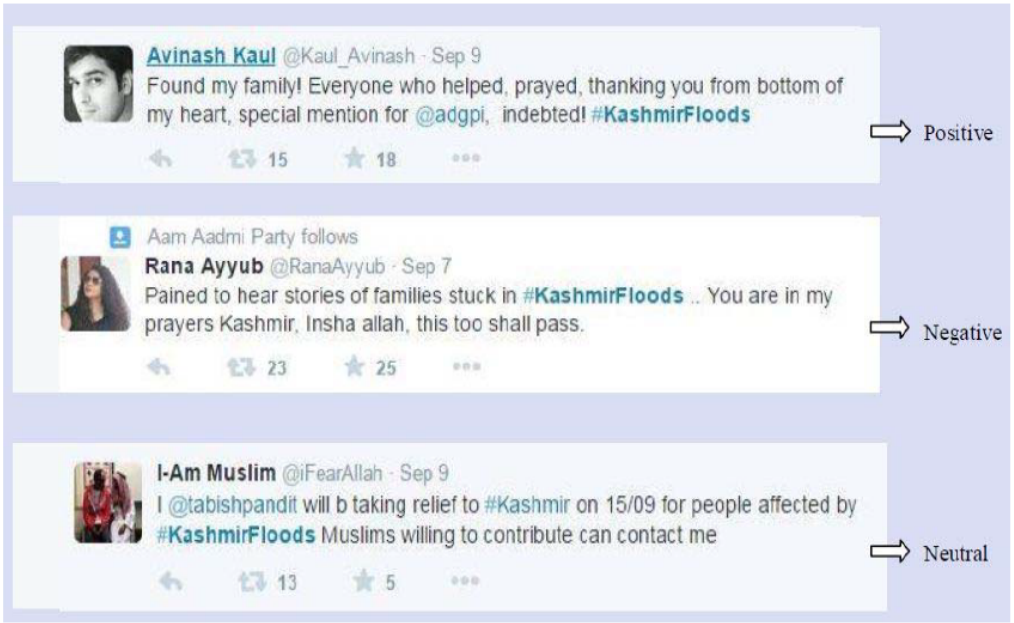

Figura 1.3: Tweets con la corrispondente categoria

In [10] viene utilizzata una implementazione del μ Model che consiste dei seguenti passaggi per il processamento dei dati: rimozione delle stopwords:

termini come "its, a, out, here, in and the", rimozione delle stemwords, uno stemmer rimuove i suffissi e a volte i prefissi lasciando solo la radice della parola, correzione degli errori di spelling, dove possibile tenendo in considerazione le convenzioni accettate per il micro text, determinazione dell'argomento, confrontando le parole chiave rimaste dai passaggi precedenti con una lista di termini a proposito del clima, già categorizzati. Tale lista è stata creata manualmente dopo l'analisi dell'archivio dei Tweet e nello svolgimento delle varie fasi, viene modificata. Tutto questo viene fatto tramite N-gram e Jaccard o simili. Inclusi nei Tweet ci sono anche quelli provenienti delle organizzazioni ufficiali. Questi sono separati dai Tweet definiti "pubblici" e sono trattati come"ground truth" contro questi, infatti, l'algoritmo verrà testato. Dopo aver separato i Tweet nelle categorie "ground truth" e "pubblici", i Tweet pubblici vengono processati da diversi algoritmi.

1.1.3 Conclusioni

In conclusione al lavoro[2] il loro modello risulta funzionale e abbastanza generico per poter essere applicato ad altri contesti oltre agli eventi sismici. Sottolineano infine una problematica, i sensori sociali non sono distribuiti in modo uniforme sul territorio e sono raggruppati nelle città e la loro efficacia è molto ridotta nelle ore notturne. In[7] è stato possibile creare una mappa di intensità dell'evento in studio (l'uragano Sandy) figura **??**fig:L4), combinando due aspetti dell'informazione: la sua geolocalizzazione e l'intenzionalità dell'utente all'evacuazione combinando i dati di geolocalizzazione e dati di intenzione di evacuazione. La mappa presenta l'importante caratteristica di essere aggiornabile in tempo reale (dinamica), rendendo possibile la valutazione dell'entità dell'evento e della conseguente evacuazione. Si è stati inoltre in grado di valutare lo stato d'animo generale comparando i Tweet raccolti in un luogo con la distanza dall'evento potenzialmente dannoso. I loro risultati mostrano come il SM sia uno strumento in grado di dare una maggior consapevolezza della situazione durante gli eventi disastrosi nonché una maggior capacità di comprendere il sentimento delle persone. In[8] hanno presenta-

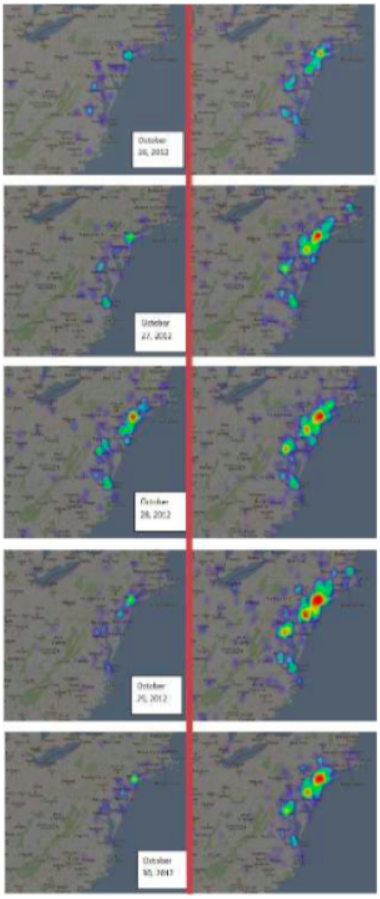

Figura 1.4: Mappa d'intensità degli utenti Twitter con intento di evacuare, La colonna di sinistra mostra gli utenti intenzionati ad evacuare mentre quella di destra intenzionati a rimanere

to nel dominio "situazioni di calamità", un sistema per il rilevamento dei sentimenti che dà informazioni utili che potranno essere usate da agenzie governative e non per gestire situazioni drammatiche in modo migliore. In[19] la ricerca ha mostrato come in condizioni climatiche eccezionali (come i temporali) è possibile avere un resoconto delle condizioni climatiche analizzando Twitter. Sia i terremoti che i temporali sono eventi a proposito dei quali le persone pubblicano su Twitter. Questa ricerca tuttavia mostra che in condizioni climatiche normali, come una giornata soleggiata, i Tweet pubblicati non sono così comuni e così descrittivi. Da questo concludono che l'ana-

lisi automatica di Twitter può produrre risultati accurati (oltre il 90%) in condizioni climatiche drammatiche o quando avviene un cambiamento nelle condizioni metereologiche.

Altri studi[18] mettono in evidenza come sia possibile rilevare le "bursty areas", zone in cui il numero di Tweets georeferenziati è particolarmente elevata. Lo studio pone l'accento su due problemi: alcune zone non vengono catalogate come "bursty" perchè molti dei Tweet non sono georeferenziati, mentre in altre, nonostante l'elevata mole di dati, il lasso di tempo tra un rilevamento e l'altro è troppo lungo per considerare "calda" l'area interessata.

1.2 Analisi qualitativa

Ciò che abbiamo esposto fino ad ora non tiene molto in considerazione il testo contenuto nei Tweet, per farlo bisogna parlare della sentiment analysis (SA) che è l'atto di processare il testo e di estrapolarne informazioni. La SA, come descritto in[3], può essere fatta su tre livelli. Il primo livello di analisi è il documento, da questo è dedotto il sentimento prevalente, che sia esso negativo, positivo o neutrale, di tutto il messaggio. Il secondo livello è la frase e permette di determinare il sentimento generale per ogni frase presente nel documento. Il terzo sono entità e aspetto, tale livello rappresenta un'analisi fine nella quale l'obbiettivo è rivelare i sentimenti di ogni entità e/o i loro vari aspetti.

In letteratura i lavori riguardanti l'analisi testuale dei sentimenti hanno sempre seguito due approcci: il primo, calcolare il sentimento dell'intero documento come la media dell'orientamento semantico (polarità) delle parole e delle frasi; questo richiede l'uso di un dizionario di parole con relativa polarità e l'indicazione se una parola specifica appartiene alla classe delle parole positive o a quelle negative. Il secondo metodo è basato sulle tecniche di Machine Learning che trattano il problema della SA come una questione di classificazione del testo[16]. Per costruire i classificatori sono usati diversi metodi di Machine Learnings e sono testati su dataset disponibili, usando

diverse caratteristiche come la presenza di un termine, la sua frequenza, unigrams, bigrams, n-grams, Part Of Speech-POS tags, micro blogging etc.[5]. Le tecniche di Machine learning sono classificate come supervised, unsupervised e semi-supervised. Nella tecniche di Supervised Learning ci sono due set di dati: training e test. I dati training sono usati per insegnare al classificatore a sistemare i dati nelle rispettive classi mentre i dati test sono usati per valutare le performance del classificatore.

Le tecniche Unsupervised Learning sono basate sul lessico, qui non è fatto alcun training per etichettare i dati. Si usa invece un database lessicale per calcolare il punteggio di una parola, la sua orientazione, confrontando le caratteristiche di un dato testo con il database. Semi Supervised Learning utilizza una combinazione di Machine Learning e approcci basati sul lessico. Si ritiene che la SA tramite Twitter sia complessa a causa della lunghezza massima di 140 caratteri per messaggio, del linguaggio informale usato dagli utenti di questo social network, dalla presenza di slang e di emoticons per esprimere opinioni etc. Esistono tre principali criteri di studio della SA in Twitter e sono basati su: parole chiave, regole linguistiche e classificazione. Sono stati fatti molti lavori volti a migliorare questi metodi. Tra questi tre il metodo basato sulla classificazione è ritenuto il migliore, quando si utilizza Twitter, perché le parole chiave e le regole linguistiche non si possono applicare a tutti i linguaggi.

Vediamo ora lo stato dell'arte nel campo del SA di Twitter: Go et al. [6] esplorano la possibilità di combinare le varie caratteristiche di n-grams con le Part of Speech Tags (POS) per le caratteristiche dei dati di "training". Sviluppano un framework per implementare la combinazione dei diversi classificatori di Machine Learning come Naive Bayes, Maximum Entropy (MaxEnt), Support Vector Machines (SVM) con vari features sets come unigram e bigram per determinare i sentimenti dei Tweet. Le Emoticons ":-)" sono utilizzate per identificare i sentimenti positivi e ":-(" per quelli negativi; usano le twitter API per la raccolta dati. La miglior accuratezza è dell' 83% e la ottengono utilizzando MaxEnt e come features unigram e bigram. Barbosa

and Feng[3] propongo, invece, di usare le caratteristiche speciali di twitter come emoticons, hash tags etc. al posto degli n-grams come suggerito da Go et al.[6] e questo dovrebbe aumentare le performance di classificazione. Suggeriscono due passaggi per rilevare i sentimenti: il primo è rilevare la soggettività per separare l'insieme di dati in due classi, soggettivi e oggettivi; il secondo è la rilevazione della polarità per classificare le frasi in positive e negative. I dati di "training" sono raccolti da tre diversi siti, Twends, Twitter Sentiment, TweetFeel, per rilevare in tempo reale i sentimenti dei Tweet. Sono state utilizzate due caratteristiche per classificare i Tweet: meta features dei Tweet e tag POS che tendono ad esprimere più sentimenti, come ad esempio aggettivi, esclamazioni e una mappa di parole. La seconda è la sintassi, emoticons, punteggiatura, lettere maiuscole, link, hash tag, RT(reTweet) ecc.

É stato creato un set usando entrambe le caratteristiche sopraelencate ed è stato testato usando tecniche di Machine Learning presenti in WEKA. Al di fuori di WEKA, SVM ha ottenuto la miglior accuratezza, 81.9% per la rilevazione della soggettività e 81.3% per la polarità. I lavori sopracitati non valutano però i sentimenti neutrali. Peak and Paroubek[15] mostrano come creare automaticamente un corpo per l'analisi dei sentimenti da twitter e costruiscono un classificatore in grado di determinare al meglio sentimenti positivi, negativi e neutri. La stessa procedura è stata utilizzata da[7] per raccogliere sentimenti positivi e negativi, mentre i post oggettivi, usati come riferimento, sono presi da account Twitter di giornali e riviste. Il cassificatore Naive Bayes è stato istruito da una combinazioni di n-grams e POS a rilevare i sentimenti dei Tweet. Pak e Paroubek raggiungono i migliori risultati, quindi la miglior accuratezza, con i bigrams. Molti ricercatori hanno esplorato l'area delle caratteristiche dell'ingegneria per migliorare le performance della classificazione dei sentimenti dei Tweet.

Agarwal et al.[1] hanno sperimentato tre modelli per la classificazione: baseline unigram model, feature based model e tree kernel based. Nel feature based model sono state proposte un totale di 50 caratteristiche, tra queste

le più importanti sono quelle che combinano la polarità delle parole e i POS tags. Mentre nel modello tree kernel based è stata disegnata una nuova rappresentazione ad albero per unire più categorie di caratteristiche ed è stato usato un partial tree-PT kernel per calcolare le somiglianze fra i due alberi[13]. Il risultato dell'esperimento sopracitato mostra che il modello feature based raggiunge un'accuratezza simile al modello unigram mentre i modelli tree kernel based hanno prestazoni significativamente migliori.

Kouloumpis et al.[9] confrontano diverse caratteristiche, incluse le n-gram features, lexicon features, POS features, e micro blogging features, ad esempio: la presenza di emoticon, abbreviazioni e rafforzativi per la classificazione dei sentimenti in Twitter. Concludono che le caratteristiche del micro blogging sono le più utili per la SM in Twitter mentre le caratteristiche POS possono non essere così utili e le lexicon features unite alle caratteristiche del micro blogging possono risultare invece utili. Tra le varie applicazioni della SA Twitter, Tumasjan et al.[19] hanno studiato la possibilità di analizzare le attività di Twitter durante le elezioni per prevedere i risultati. Hanno collezionato 104,003 Tweet che citano partiti e politici prima e durante le elezioni federali tedesche del 2009. Questi Tweet sono stati analizzati per rilevare i sentimenti utilizzando un software per l'analisi del testo chiamato LIWC2007(Linguistic Inquiry and Word Count)[14] .

I risultati hanno mostrato che Twitter è stato effettivamente utilizzato come piattaforma per il dialogo politico e anche il numero di Tweet riferiti ad un particolare partito è accurato quasi quanto i tradizionali sondaggi elettorali. Recentemente sono state svolte molte ricerche per valutare la possibilità di usare l'analisi dei social media per la gestione di eventi catastrofici e la diffusione delle informazioni in modo da aumentare la consapevolezza di ciò che accade [4] . Ci sono però pochi lavori per quanto riguarda l'analisi dei sentimento(SA) durante i momenti di crisi come terremoti, esplosioni e uragani [16, 17].

1.3 Gestione dell'emergenza tramite Social Media

Lo studio della letteratura in merito all'utilizzo dei Social Media per l'analisi in tempo reale degli eventi ascrivibili alla categoria di emergenza, ha messo in luce quanta rilevanza abbia questo metodo nel contesto sociale presente e futuro. I SM sono in continua evoluzione, ed è sempre maggiore e capillare la loro diffusione e il numero di utenti che quotidianamente fruisce del servizio. Conseguentemente, aumenta ogni giorno la mole di dati prodotta. Il 48 rapporto Censis rivela che, solo in Italia, gli utenti attivi su Facebook, con un'età compresa tra i 35 e i 45 anni anni, dal 2009 al 2014, sono aumentati del 153% mentre gli over 55 persino del 405%. Facebook continua ad essere il social network di maggior utilizzo, ma tutti risultano in crescita.

Esistono numerosi strumenti già certificati e distribuiti, o in fase di perfezionamento, che basano il loro funzionamento sull'analisi dei dati provenienti dai SM. Fra questi, vale la pena ricordarne alcuni, che mostrano la rilevanza applicativa dell'analisi dei SM. SMEM (Social Media Emergency Manager) si propone di promuovere l'uso di un hashtag, #smem, per segnalare le emergenze o altri eventi di interesse sociale. Alert4All mira a migliorare l'efficacia dei messaggi di avviso e di comunicazione con la popolazione in caso di calamità e si concentra sul ruolo di SM nelle comunicazioni di emergenza.

Il progetto SMART-C raccoglie e integra i dati provenienti da fonti diverse come i social network, blog, comunicazioni telefoniche di rete fissa, SMS, MMS. Il sistema SHIELD è stato sviluppato per sfruttare i dispositivi mobili, al fine di ridurre i tempi di risposta e di soccorso per le vittime della criminalità nei campus universitari degli Stati Uniti. Diversi studi si sono rilevati efficaci nel predire la diffusione dell'influenza H1N1 in Gran Bretagna (95% di accuratezza)[20] o nello studio di terremoti e uragani, come ad esempio il già citato uragano Sandy.

Naturalmente, ogni studio mette in evidenza nuovi limiti e traccia la stra-

da per rendere sempre più efficace un sistema che già si pone come un valido e innovativo strumento di previsione e gestione delle emergenze. Viene rilevato come resti ancora problematico un processamento preciso dei dati, che comprenda tutte le fonti necessarie, ma sia in grado di non tenere in considerazione i falsi positivi. Un primo passo, quindi, potrebbe essere l'impostazione automatica di parametri di selezione dei dati. Inoltre, occorre comunque ottenere un riscontro basato sulle informazioni di canali ufficiali [20].

Un'altra limitazione piuttosto evidente riguarda la georeferenziazione dei dati, non sempre presente o non sempre accessibile, spesso non accurata, che rende, di fatto, inutile il dato al fine di uno studio. Correlato a questo, occorre prendere in considerazione il fatto che i "sensori sociali", ovvero gli utenti che producono un messaggio, non hanno una distribuzione uniforme sul territorio, il che rende difficile l'analisi in zone scarsamente popolate[2]. Nello stesso studio viene citato il problema, non affrontato, della sicurezza. Occorre la possibilità di verificare quanto il dato sia attendibile, in modo da eliminare possibili dati fittizi generati da una persona o da un gruppo di persone.

Articolo	Evento Rilevato	API	SM	Algoritmo	KeyWords
[20]	Influenza Pandemica	Non Specificato	Twitter	flu-score, numero di marker fratto il numero totale di marker	Fever, Temperature, Sore throat, Infection, Headache
[7]	Analisi delle informazioni di evacuazione durante l'uragano Sandy	Search API	Twitter	Latent Semantic Indexing, Granger causality analysis	Evacuate, Prepare, Flood, Hurricane, Storm, Tornado, Sandy, Rain, Water, Food, Emergency, Leave
[8]	L'inondazione del Kashmir (autunno 2014)	Search API	Twitter	Unigram features using Naïve Bayes.	#kashmirfloods #jammufloods.
[18]	Situazioni di Emergenze, in particolare la pioggia	Streaming API	Twitter	(ϵ,τ)-density-based spatiotemporal clustering algorithm	Rain
[2]	Terremoti	Streaming API	Twitter	Calcolo della frequenza dei dati in un intervallo di tempo	Scossa, Terremoto
[10]	Condizioni Meteriologiche	Search API	Twitter	μ Model	#ptaweather , #pta @PretoriaZA #pretoria, #weather
[21]	Situazioni di emergenza cittadine	Weibo API	Weibo	Similarità del coseno, GIS based visualization	Fire (in cinese)

Tabella 1.1: Tabella Riassuntiva

Capitolo 2

Twitter

2.1 Origini

Twitter è una popolare piattaforma di social networking e micro-blogging che ha fatto la sua prima apparizione il 21 marzo del 2006. Twitter nasce dall'idea di alcune giovani menti americane che vedono i loro primi sforzi prendere forma in Odeo Inc., società nata con l'intento di sviluppare una piattaforma di podcasting il cui sviluppo fu però interrotto, prematuramente, dal rilascio da parte di Apple della propria piattaforma di podcasting interna.

Decisero allora di reinventare il proprio prodotto, così si diffuse l'idea di un servizio che permettesse di condividere, in tempo reale, cosa e dove stesse accadendo qualcosa. L'idea iniziale prendeva spunto dagli SMS di gruppo con i quali una persona poteva condividere programmi per la serata o avvenimenti, con i propri contatti più stretti.

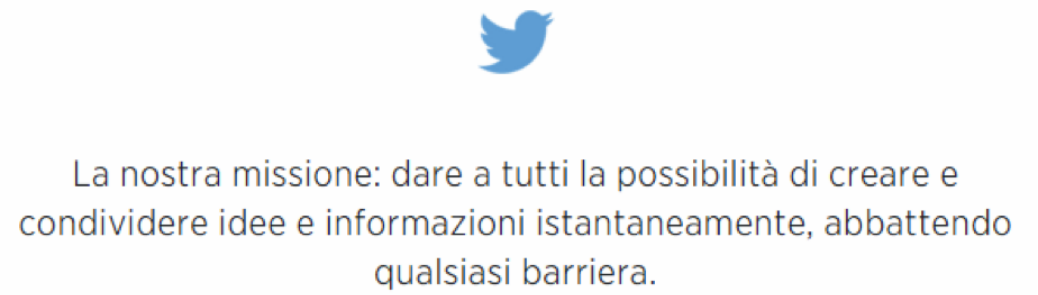

Figura 2.1: Il motto di Twitter [24]

Da allora la piattaforma ha continuato a evolversi ed espandersi, fino a diventare un punto di riferimento nel mondo dei social network nonchè una risorsa per ottenere notizie in tempo reale. Twitter cominciò ad essere notata da cronache locali e specializzate in tecnologie grazie al fatto di essere stata usata, nei primi giorni di vita, per dare notizia di un terremoto a San Francisco. Nel primo anno di vita il social network di San Francisco, riesce a moltiplicare per tre il numero dei Tweet, passando dagli iniziali 20mila a 60mila. Nel 2009 arriva la notorietà su scala planetaria, grazie ai disordini in Iran e alla cosiddetta primavera araba, avvenimenti che trovarono in Twitter uno dei mezzi più usati per seguire i fatti in diretta e ascoltare le voci di testimoni e protagonisti. Infine anche il mondo della politica e delle istituzioni capisce le potenzialità dello strumento, e inizia a farne un uso massiccio per la comunicazione e le campagne elettorali.

I numeri oggi sono cresciuti in modo esponenziale: 218 milioni di utenti, 100 milioni quelli giornalmente attivi, non hanno soltanto la possibilità di postare messaggi da 140 caratteri, ma anche foto e clip audio e video: è del 2012 l'acquisizione dell'applicazione Vine, utile proprio per questi scopi. E la prossima frontiera potrebbe essere quella di dare vita a una chat sull'esempio di Whatsapp[23].

2.2 Twitter APIs

2.2.1 API brevi cenni

Per API (Application Programming Interface) si intende il complesso di operazioni, tipi e caratteristiche funzionali associate ad uno specifico programma ed esposte agli sviluppatori. Le API permettono di estendere con semplicità le funzionalità di un programma, offrendo la possibilità di interazione con i componenti del programma stesso. Le API possono variare da specifiche librerie collegate a routine, strutture dati e classi di oggetti del programma, a chiamate remote (ad esempio per servizi REST) esposte agli

Figura 2.2: Diffusione e utilizzo di Twitter al 31/12/2015 [24]

utilizzatori delle API stesse. Due sono le tipologie di API principali fornite da Twitter: le REST APIs e le STREAMING APIs.

2.2.2 REST API

Le API REST forniscono accesso programmatico alla lettura e alla scrittura dei dati di Twitter: autori e applicazioni con cui sono pubblicati i Tweet vengono identificati tramite OAuth[1] mentre le risposte sono salvate in strutture JSON[2]

Limitazioni delle REST APIs Le API REST sono limitate nel numero di richieste disponibili al singolo utente, o meglio, a livello di singolo token

[1]Protocollo open che permette l'autenticazione sicura tramite applicativi mobili, desktop e siti web. Per una disamina comprensiva del protocollo e del relativo framework, si veda http://oauth.net/

[2]Javascript Object Notation, formato per lo scambio dati, facilmente intellegibile e utilizzabile sia a livello umano che di macchina preposta a generarne e analizzarne la sintassi.

Web. Il social network dichiara 18 milioni di contatti complessivi, ma sono di meno gli utenti attivi stimabili

@SOLE24ORE
18 NOVEMBRE 2015

Su Twitter otto milioni di italiani

«Vi vorrei segnalare questo numero, per nulla banale: la reach complessiva di Twitter in Italia è di 18 milioni di utenti unici al mese». Eccolo il primo numero "ufficiale" di Twitter nel nostro Paese.

È stato lo stesso ceo di Twitter in Italia, Salvatore Ippolito, a comunicarlo durante «PoweredByTweets, l'evento che si è tenuto ieri a Milano all'Unicredit Pavillon. Presentatrice Geppi Cucciari, e "incursione" dell'ospite a sorpresa Fiorello, la serata è stata l'occasione per fare il punto su Twitter in Italia. In fondo una classica serata di "upfront" come si direbbe in gergo televisivo. «È un termine – replica Ippolito al Sole 24 Ore – in cui sinceramente non mi riconosco. Volevamo invece fare il punto dopo i primi due anni di attività nel Paese, spiegando chi siamo e quello che facciamo».

Attenzione: i 18 milioni di utenti comprendono gli utenti attivi ma anche quelli raggiunti – e quindi contatti per i banner pubblicitari – attraverso tutte le App che con Twitter hanno accordi, formando un network. È invece impossibile scucire un numero ufficiale sugli utenti iscritti in Italia (nel mondo sono 320 milioni). Un tentativo si può azzardare facendo un parallelo con la situazione internazionale. Se infatti a livello globale ci sono 320 milioni di iscritti e oltre 700 milioni nella "total audience" di Twitter, i 18 milioni della total audience in Italia dovrebbero condurre a poco più di 8 milioni di iscritti. Il numero è in linea peraltro con le stime che circolano.

Quel che invece è certo è che il social ha una situazione finanziaria a livello globale con ricavi in crescita oltre le attese (a 569 milioni di dollari, +58%, nel Q3 2015) ma utenti in crescita ben sotto le aspettative: 4 milioni in più rispetto al trimestre precedente e sotto gli 8 milioni attesi dagli analisti.

In questo quadro negli ultimi tempi si sono moltiplicate le analisi su quella che a molti è apparsa come una ineludibile virata verso il basso per un social che a livello globale subisce evidentemente il confronto con Facebook. Anche in Italia? «Io posso solo dire – replica Ippolito – che in Italia stiamo andando benissimo. E che abbiamo un'utenza molto pregiata. I numeri lo testimoniano».

Durante la serata ne sono stati citati molti. Per esempio che il 67% di utenti Twitter lo usa ogni giorno e il 46% lo usa più volte al giorno. Oppure che il 52% di utenti "cinguetta" appena sveglio e il 54% lo fa come ultima attività prima di mettersi a dormire. Inoltre c'è un 80% che usa Twitter da mobile, oppure un 83% che segue su Twitter brand o aziende. E poi c'è l'interazione molto forte con la regina del mercato pubblicitario, la Tv, visto che 1 utente su 2 in Italia usa Twitter mentre guarda al tv e 9 su 10 twittano sullo show che stanno guardando. «Sono numeri – dice il ceo di Twitter Italia – che confermano la nostra validità anche come partner pubblicitario». Non a caso ieri sono state anche premiate le migliori campagne con dei "Twitter Awards". Premiate Ford con la Campagna «SanValentino; la Juventus con la campagna «SeAndiamoInFinale; Mercedes per la migliore campagna su Periscope («FindTheSuv) e Telecom per la campagna che ha fatto un utilizzo più innovativo della piattaforma con il «TimStadium durante il derby di Milano.

A. Bio.

Figura 2.3: Aritoclo del Sole24ore del 15/11/2015

di accesso. Temporalmente, le soglie limite vengono suddivise a livello di intervalli di 15 minuti, oltre i quali il numero di richieste effettuate viene resettato; inoltre, tutti gli endpoint richiedono autenticazione al momento della comunicazione per impedire richieste non autorizzate e tenere traccia delle soglie limite.

Per le richieste di tipo GET vi sono due limitazioni iniziali, 15 chiamate ogni 15 minuti e successivamente 180 chiamate in 15 minuti. Similmente, la ricerca è limitata a 180 query ogni 15 minuti. Per indagare lo stato delle richieste della propria applicazione rispetto ai limiti di cui sopra, è possibile ispezionare gli header HTTP delle richieste stesse, sia che si utilizzi autenticazione a livello di applicazione che a livello di singolo utente. In particolare:

- X-Rate-Limit-Limit - specifica il tetto massimo per la specifica richiesta
- X-Rate-Limit-Remaining - specifica il numero di richieste rimanenti nella finestra di 15 minuti
- X-Rate-Limit-Reset - specifica il tempo rimanente prima dello scadere della finestra di 15 minuti, espresso in Epoch Time[3]

[3]Numero di secondi trascorsi dal 01/01/1970 00:00:00 Coordinated Time Universal (UTC)

Le richieste GET e POST sono limitate le une a livello di utente e applicazione (il conteggio è legato anche all'applicazione dalla quale sono state inoltrare le richieste) mentre le altre a livello di utente. In questo modo, si punta a limitare un'eccessiva abbondanza di richieste POST da applicativi differenti, onde evitare comportamenti inconsistenti all'atto della risposta. In caso i limiti siano superati, la API restituisce il codice HTTP 429 "Too Many Requests", per indagare i limiti si può sfruttare la richiesta GET application/rate_limit_status [4]. Il superamento delle soglie limite può comportare, in caso di cattiva gestione delle risorse a disposizione, fenomeni di blacklisting di account o applicazioni, con conseguente insensibilità da parte delle API a successive richieste. Per evitare il superamento delle soglie, è possibile mettere in atto svariate best practices, quali:

- Utilizzare metodologia di caching delle risposte fornite dalle API, che permette un maggior afflusso di traffico verso la cache piuttosto che verso l'API stessa
- Gestire prioritariamente le richieste provenienti da utenti attivi, o che hanno effettuato login recentemente
- Mantenere traccia di quali query producano risultati soddisfacenti, filtrando le query che generano molto traffico a fronte di scarsi benefici

2.2.3 Search API

All'interno delle API REST, la principale è sicuramente la Search API, che permette di effettuare query sul set di Tweet pubblicati nel corso degli ultimi sette giorni; a differenza della Streaming API, esposta in dettaglio più avanti nella trattazione, questa API di ricerca fa della rilevanza, più che della completezza, il suo focus primario. Fondamentale è sfruttare la Search API senza superare i limiti ad essa imposti, dedicando tempo al tuning delle query di ricerca e all'implementazione di tecniche di caching dei risultati attesi. Le

[4] Un elenco comprensivo dei limiti di richieste nel tempo è disponibile sotto forma di tabella all'indirizzo https://dev.twitter.com/rest/public/rate-limits

risposte derivanti da query, effettuate tramite questa API, possono mancare di alcuni Tweet o utenti, a favore di una migliore corrispondenza dei risultati a quanto cercato.

Dato il presupposto di cui sopra, l'accurata costruzione delle query di ricerca risulta chiave dell'ottenimento di risultati significativi. Molti operatori compongono lo scheletro delle query di ricerca e possono essere suddivisi per fasce di complessità crescenti.

Ricerca base Nel suo utilizzo più immediato, la Search API permette di cercare testo libero all'interno del Tweet: la ricerca supporta i comuni operatori AND (operatore base della ricerca, che non necessita di essere esplicitato) e OR (che deve essere esplicitato) per query il cui risultato prevede la presenza di tutte o alcune delle parole cercate. Per meglio concentrare lo sforzo di ricerca è possibile inserire nella query una lista di parole comprese tra doppi apici "" e sfruttare la Search API per ricercare una specifica frase (e non le singole parole); è inoltre utile sfruttare il segno "-" per scartare i risultati contenenti una parola non ritenuta di interesse per la ricerca. Infine, ma non meno importante, è la ricerca attraverso il carattere di hashtag "#", che riporta i risultati contenenti l'argomento di interesse.

Ricerca per utente Le query possono interessare specificatamente un utente o le sue interazioni. In particolare, è possibile sfruttare il selettore "@" per indagare Tweet che menzionano l'account cercato, "from:" e "to:" per la ricerca di Tweet inviati da e in risposta all'account specificato, senza contare le menzioni spontanee di terzi, ed infine "list:" con specifica di "account/-lista" per i Tweet inviati da account, inseriti nella lista creata dall'account ricercato.

Filtri espliciti Uno degli operatori utilizzabili all'atto della ricerca è "filter:" che, associato ad una parola chiave, permette di restringere i risultati ai soli Tweet contenenti specifici elementi propri della struttura di Twitter. L'operatore "filter:" non viene infatti utilizzato per la ricerca testuale, ma per

estrapolare dai risultati quelli contenenti lo specifico media di interesse (foto, video, link) o di escludere contenuti potenzialmente sensibili ("filter:safe").

Ricerca temporale La ricerca può essere coadiuvata dagli operatori temporali "since:" e "until:" che selezionano i risultati prodotti sino o da una certa data in formato yyyy-mm-dd.

Ricerca tramite sentiment analysis Alcuni degli operatori più avanzati riguardano la possibilità di costruire una query che filtri i risultati secondo l'accezione positiva o negativa della frase contenuta nei Tweet, o in base a specifici segni di punteggiatura: in questo caso vengono allegati alla stringa di ricerca i simboli di punteggiatura richiesti "?", "!", etc. oppure smiles stilizzati che riproducano l'accezione desiderata (":)" per la positiva, ":(" per la negativa).

Alla stringa componente la query possono poi essere aggiunti ulteriori parametri quali: l'ottenimento dei soli Tweet più popolari o di tutti quelli recenti, la lingua in cui sono composti i Tweet o latitudine e longitudine di ricerca dei Tweet geolocalizzati.

Tweets by Place Una delle estensioni principali della Search API è la possibilità fornita dall'operatore "place" che implementa la ricerca geolocalizzata attraverso l'ID codificato di un luogo, senza la necessità di utilizzare la sua latitudine e longitudine. Vari sono i servizi che forniscono ID di luoghi ricercabili tramite operatore "place", a partire dal servizio interno di Twitter fino a GoWalla o TomTom.

2.2.4 Streaming API

Le Streaming APIs implementano un accesso a bassa latenza allo stream globale dei dati di Twitter. Forniscono push di nuovi Tweet al loro accadere senza l'overhead associato, a differenza dell'accesso tramite REST APIs e della gestione delle sue richieste.

Differenze tra Streaming e REST La connessione alle Streaming APIs richiede una connessione HTTP persistente al servizio, a differenza di quanto avviene attraverso le REST APIs in cui ogni richiesta viene processata in modo indipendente e ne viene poi fornito il risultato. Per questo, è solitamente consigliabile mantenere nella propria applicazione un processo separato che gestisca la connessione di tipo streaming, permettendo parsing, filter e aggregazioni necessarie sul momento. L'implementazione delle richieste di tipo streaming è più complessa ma fornisce la possibilità di visualizzare lo stream di Tweet in real-time con una lista più fedele delle interazioni che hanno avuto luogo su Twitter.

Come nelle REST API, è possibile raffinare la ricerca per evitare di incorrere nelle limitazioni, ma anche per evitare di ricevere messaggi ai quali non si è interessati. I parametri utilizzabili si possono riunire in gruppi, come specificato nella documentazione a disposizione degli sviluppatori fornita da Twitter, e sono i seguenti:

Lingua Impostare questo parametro permette la visualizzazione solo dei Tweet scritti nella lingua di interesse scelta. Per esempio, *language=en* restituirà esclusivamente i Tweet scritti in lingua Inglese.

Follow Per usufruire di questo filtro si crea una lista, separata da virgole, di ID utenti, così verranno consegnati solo i Tweet appartenenti a questi utenti. Per ogni utente così specificato lo stream conterrà:

- Tweet creati dall'utente
- Tweet che sono stati oggetto di reTweet da parte dell'utente
- Risposte ad ogni Tweet creato dall'utente
- Tweet creati dall'utente che sono stati oggetti di reTweet
- Risposte manuali che sono state create senza aver fatto click sul pulsante adibito alla risposta

Lo stream non conterrà:

- Tweet in cui è menzionato l'utente
- ReTweets manuali
- Tweet di utenti protetti

Track Con questo filtro è possibile fornire una lista, separata da virgole, di frasi che saranno usate per decidere quali Tweet saranno recapitati all'applicazione. Una frase può essere composta da una o più parole e ogni frase, per combaciare coi criteri di selezione, dovrà avere ogni parola presente nel Tweet mentre l'ordine è indifferente. Secondo questo modello le virgole possono essere intese come operatore logico "OR", gli spazi, invece, come "AND" (ad es. "the twitter" è un AND e "the, twitter" è un OR). La tabella 2.1 mostra alcuni esempi (fonte: dev.twitter.com/streaming/overview/request-parameters).

Località Questo filtro fornisce una lista di punti, formati da latitudine e longitudine, che formeranno un perimetro con cui verranno filtrati i Tweet. In questo modo solo i Tweet con le informazioni di geolocalizzazione che risiedono dentro il perimetro fornito verranno consegnate al client.

Limitazioni Come le API REST anche le Streaming API soffrono di limitazioni, la più importante è che il volume in ingresso non può superare l'1% del traffico totale di Tweet, ce ne sono altre invece che riguardano le connessioni dei client, alcune di queste sono: i client che non implementano il backoff, e tentano di ricollegarsi il prima possibile, avranno un tasso di connessioni limitato per un breve intervallo di minuti. I client con tasso di connessioni limitato riceveranno risposte HTTP 420 per tutte le richieste di connessione. I client che attivano e disattivano la connessione frequentemente, per esempio per modificare i parametri di ricerca, corrono il rischio che venga limitato il loro tasso di connessioni. Twitter non rende pubblico il

Parameter value	Will match	Will not match
Twitter	TWITTER twitter “Twitter” twitter. #twitter @twitter http://twitter.com	TwitterTracker #newtwitter
Twitter's	I like Twitter's new design	Someday I'd like to visit @Twitter's office
twitter api, twitter streaming	The Twitter API is awesome. The twitter streaming service is fast. Twitter has a streaming API	I'm new to Twitter
example.com	Someday I will visit example.com	There is no example.com/foobarbaz
example.com /foobarbaz	example.com/foobarbaz www.example.com/foobarbaz	example.com
www.example.com /foobarbaz		www.example.com /foobarbaz
example com	example.com www.example.com foo.example.com foo.example.com/bar I hope my startup isn't merely another example of a dot com boom!	

Tabella 2.1: Esempi di filtro Track

numero di tentativi di connessione che causa una limitazione della velocità, comunque c'è una soglia di tolleranza per test e sviluppo. Alcune dozzine di tentativi di connessione di tanto in tanto, non attivano il limite. Tuttavia, è necessario evitare ulteriori tentativi di connessione per pochi minuti se viene ricevuta una risposta 420 HTTP. Se il client è limitato di frequente, è possibile che venga bloccato l'accesso a Twitter all'IP per un periodo di tempo

indeterminato. In ultimo un'altra importante limitazione sta nel fatto che se vengono creati messaggi, mentre il client non è connesso, questi saranno persi, è possibile ricevere i Tweet solo dopo che la connessione è avvenuta.

Capitolo 3

Sistema di rilevamento di eventi su dati Twitter

In questo lavoro si è voluto ideare e realizzare un semplice sistema per la rilevazione di eventi nella regione Emilia Romagna tramite l'analisi di Tweet geolocalizzati nella regione di interesse. L'idea dalla quale si è partiti per l'ideazione del metodo di rilevazione è stata data dal lavoro di Avvenuti et al del 2010 [2], in questo lavoro è stato infatti osservato che l'accadere di un evento porta a un aumento del numero di messaggi, relativi a tale evento.
La nostra applicazione è strutturata in diversi moduli (mostrato graficamente nella Figura 3.1) ognuno dei quali risponde ad una necessità alla quale il sistema doveva far fronte, tali moduli sono:

- Acquisizione, colleziona i dati di Twitter
- Filtraggio, riduce il rumore di fondo eliminando Tweets non utili al nostro scopo
- Analisi Quantitativa, analizza in modo i dati raccolti
- Analisi Spaziale, analizza i dati raccolti in base alla loro geolocalizzazione
- Pagina web di controllo, si vedono i dati raccolti

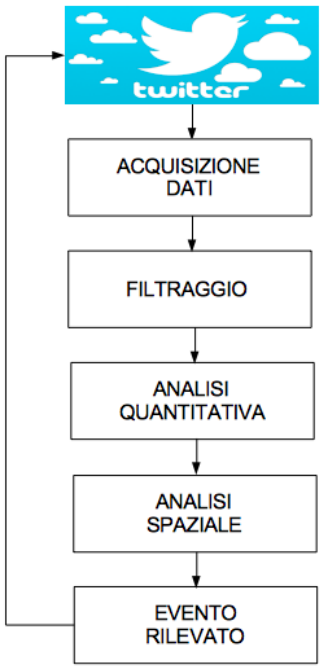

Figura 3.1: Architettura proposta

3.1 Acquisizione

Il Modulo di Acquisizione dati è il più importante in questo sistema di rilevamento, da esso derivano tutti gli altri moduli; eventuali errori commessi in questa sezione, quali la perdita di dati o l'inserimento di dati non adeguati, porteranno alla propagazione dell'errore negli altri moduli compromettendo la corretta rilevazione degli eventi.

Utilizzando le Twitter API ogni quindici minuti sono eseguite delle query, tramite le quali, vengono presi i Tweet pubblicati dalla mezzanotte del giorno stesso fino al momento in cui è stata eseguita la query. Il motivo per cui si seleziona questo intervallo di tempo è che le API non permettono la specificazione dell'orario. Insieme all'acquisizione si opera un fase di pre-filtraggio dove, rispetto a tutti i nuovi Tweet prodotti (9100 nuovi Tweet al secondo in media), si selezionano solo quelli che possono essere di nostro interesse.

Diversamente da [1, 2, 7, 8, 10, 19], nel nostro sistema non viene fatta

una query in base a parole chiave presenti nei Tweet bensì in base alla loro geo-localizzazione. Questo metodo ci fornisce solo i Tweet eseguiti in Emilia Romagna ma non necessariamente tutti, infatti vengono esclusi tutti quelli che non possiedono le coordinate di longitudine e latitudine specifiche della posizione in cui sono stati creati. Senza la geolocalizzazione l'analisi spaziale sarebbe impossibile. Esistono diverse tecniche, come l'estrapolazione del luogo in base a parole presenti nei Tweet, ma oltre ad essere difficoltose possono non dare risultati utili, in letteratura ci sono esempi di utilizzo della città in cui è stato creato l'account dell'utente che compone il Tweet ma anche questo metodo può risultare fuorviante. Tra i metodi forniti da Twitter per l'estrazione di informazioni sfruttiamo le Search API grazie alle quali il nostro modulo di acquisizione, ogni quindici minuti, fa una query di ricerca per ricevere tutti i Tweet con le informazioni di geo-localizzazione indicate prima.

Per l'implementazione di questo modulo è stato creato uno script python ed è stata usata la libreria *TwitterSearch*[1] che consente di eseguire richieste alle APIs di Twitter i Tweet restituiti, con cadenza di quindici minuti, vengono salvati in un database MySql pronti per essere filtrati e analizzati. Di seguito viene mostrato il codice che per le query alle Twitter APIs.

```
# crea un object TwitterSearchOrder
tso = TwitterSearchOrder()
#setto le chiavi per le restrizioni
tso.set_keywords(['-rt',Intervallo_tempo])
#setto la lingua
tso.set_language('it')
#setto i punti di geolocalizzazione e il raggio
tso.set_geocode(44.4990968,11.2616457,200)
#inserisco le API Key
ts = TwitterSearch(
consumer_key = '',
consumer_secret = '',
access_token = '',
access_token_secret = '')
```

[1]https://pypi.python.org/pypi/TwitterSearch/

```
15 #questo mi permette di iterare attraverso tutti i possibili /
       Tweet
16 for Tweet in ts.search_Tweets_iterable(tso):
17     #prefiltraggio e salvataggio a DB
```

3.2 Filtraggio

Utilizzare le restrizioni di geo-localizzazione per interrogare la piattaforma ci permette di raccogliere i messaggi già pre-filtrati, empiricamente ci si è accorti che nonostante questa prima selezione i dati possono essere ancora fuorvianti e per evitare questa problematica vengono fatte due scremature diverse, la prima riguarda gli account dai quali i messaggi vengono prodotti, la seconda il contenuto dei messaggi.

3.2.1 Filtraggio Account

Alcuni account come quelli ufficiali appartenenti a organizzazioni o a esercizi commerciali, sono account che portano ad avere dei picchi nel numero di Tweet che, apparentemente, potrebbero segnalare il verificarsi di un evento quando invece non lo rispecchiano.

Portiamo l'esempio dell'account *@visitparma* un account per la promozione di eventi nella città di Parma il cui numero di Tweet non segue una logica ma aumenta la sua attività per promuovere eventi o sconti per motivi economici e questo causa dei falsi allarmi. Account di attività commerciali come quella citata, possono essere silenti per molto tempo per poi ricomparire in riferimento a lavori appena compiuti o a forme di offerte relative alla propria attività allo scopo di attirare nuovi clienti massimizzando la visibilità. Questi account non sono personali ed è quindi improbabile che durante un evento siano usati per dare informazioni relative allo stesso; per evitare di falsificare le statistiche si è creata una lista di account i cui Tweet non vengono salvati nel nostro database. Il controllo sull'account viene eseguito prima del salvataggio del messaggio nella base di dati.

3.2.2 Filtraggio Contenuto

Il secondo metodo di filtraggio, simile al precedente, riguarda il contenuto dei Tweet. è stata creata una lista di parole scurrili, bestemmie e altri termini di questo genere, non consoni ad una segnalazione di un qualsiasi evento. Si pensi, infatti, ad un evento catastrofico, difficilmente un utente coinvolto da vicino perderebbe tempo ad inveire contro quanto gli sta accadendo, è altamente più probabile che si limiti alla segnalazione. Questa ipotesi ci ha portato a scartare questo genere di Tweet, poco affidabili sia ai fini statistici che alla rilevazione di un evento.

Oltre alle parole scurrili vengono scartati i Tweet che contengono parole come "buon giorno" o "buona notte" nonchè quelli in cui compare un numero di tag, superiore a quattro, riferiti ad altri utenti. Per motivare questa scelta viene portato l'esempio dell'utente *@vadabio*, questo è un utente molto attivo, nel solo mese di Gennaio ha composto più di 1000 Tweet, ciò lo porta ad essere tra i migliori possibili sensori che si trovano nella zona di interesse essendo un utente abituale della piattaforma. Anche questo utente, come *@visitparma*, è stato causa di un picco ingiustificato dei Tweet pubblicati.

Prendiamo ad esempio un messaggio che questo utente è solito comporre, il cui contenuto può variare leggermente di giorno in giorno (Fig. 3.2). Questo messaggio è evidentemente la risposta o il proseguimento di una

Figura 3.2: Tweet di esempio di @vandabio

conversazione, tale fatto lo rende inutile ai nostri fini.

La grande differenza fra il primo e il secondo filtro sono i momenti in cui essi vengono applicati; il primo viene eseguito prima del salvataggio sul

database mentre il secondo viene messo in atto subito dopo, quando il modulo per l'acquisizione è in standby.

3.3 Analisi Quantitativa

Come già detto il verificarsi di un evento viene rilevato osservando una crescita inaspettata del numero di Tweet rispetto alla normalità. La problematica di questo modulo è decretare quale sia il numero di messaggi critico tale da diventare indice dell'accadere di un evento. Nella nostra analisi ricorriamo all'outlier detection come indice di rilevamento di un evento straordinario, un outlier viene definito come un valore anomalo rispetto all'insieme delle osservazioni che stiamo considerando.

In generale la presenza di outlier può invalidare i risultati di un'analisi dei dati quindi è opportuno ricorrere a due possibili soluzioni del problema: usare metodi di analisi "robusta" (es. mediana), oppure identificare gli outlier (outlier detection) per sottoporli a trattamento, come la sostituzione del dato con "missing", se si sospetta un errore di data entry, o rimozione di record, se si sospetta che si tratti di un valore valido ma "raro". In ambito univariato esistono varie procedure per identificare gli outlier come la 3σ rule e Hampel Identifier, queste due modalità hanno bisogno di un ipotesi di simmetria distributiva che non viene garantita nella nostra analisi, e inoltre corriamo il rischio di masking per la 3σ rule, cioè un'osservazione outlying è erroneamente ritenuta "non-outlying", e di swamping per l'Hampel identifier, un'osservazione "non-outlying" è erroneamente ritenuta "outlying". Quindi avendo necessità di considerare un'ipotetica asimmetria distributiva ci avvaliamo della procedura di Box-plot rule che ne tiene conto.

La Box-plot rule si avvale dello strumento grafico Box-plot che per essere rappresentato, attraverso i dati a disposizione, calcola mediana, primo e terzo quartile; tale regola ritiene x_i un outlier se vale una di queste condizioni:

$$x_i > min[Q_3 + 1.5IQR, max\{x_i = 1. \ldots, n\}]$$

$$x_i < max[Q_3 - 1.5IQR, min\{x_i = 1. \ldots, n\}]$$

dove $IQR = Q_3 - Q_1$ è detto range interquartile. Quindi è outlier l'osservazione che si trova fuori dai baffi.

Graficamente è possibile confrontare il Box-plot con una distribuzione campanulare dove IQR corrisponde alla parte centrale della campana (50%), i baffi superiore e inferiore rappresentano le osservazioni nelle code (rispettivamente 24.65%), le osservazioni che si trovano nella parte restante (nel 0.35% delle due code) sono ritenute outlier.

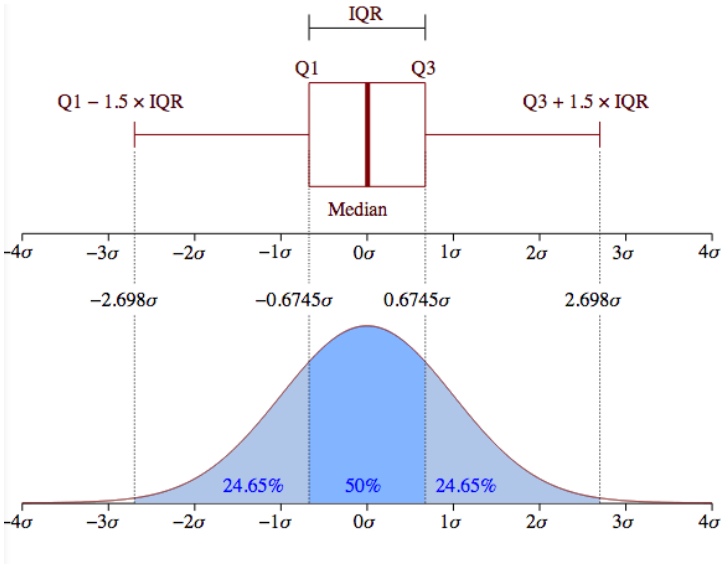

Figura 3.3: Box-plot Rule fonte: http://i.stack.imgur.com/ZN8N6.png

La sua implementazione è composta da uno script python che riceve in input un array, composto dalla somma del numero di Tweet fatti nell'intervallo di tempo in esame, e restituisce "True" se l'ultimo valore passatogli, che sarà quello corrispondente all'intervallo di nostro interesse, è un outlier. Questo è fatto con l'ausilio della libreria python *numpy*[2]. Di seguito viene mostrato il codice relativo.

```
1 def isOutLier(array):
2     a = np.array(array)
3     q1 = np.percentile(a, 25)
4     q3=np.percentile(a,75)
5     n=array[-1]
```

[2]http://www.numpy.org/

```
    IQR=q3-q1
    if n>q3+1.5*IQR:
        return True
    else:
        return False
```

Questo modulo nel nostro sistema viene richiamato due volte, la prima, con un intervallo di tempo di 15 minuti, la seconda con un intervallo di 30 minuti. Nel capitolo 4 andremo ad illustrarne il motivo.

3.4 Analisi Spaziale

L'obbiettivo di quest'analisi è appurare che il picco di Tweet misurato sia effettivamente il sintomo di un evento in atto. Per farlo, controlliamo che ci sia un picco effettivo di Tweet in una determinata zona della regione. Analizzando la distribuzione spaziale dei Tweet in una giornata (Fig. 3.4), si può notare come siano distribuiti preferenzialmente lungo l'asse della via Emilia.

Questo risultato non ci sorprende in quanto è proprio lungo questo asse che si concentra maggiormente la popolazione (Fig.3.5).

Figura 3.4: Distribuzione Tweet in una giornata

Si è deciso, quindi, di dividere i marker gps in dieci gruppi, ognuno dei quali ha il proprio centro in una delle città che mostra una grande densità abitanti:

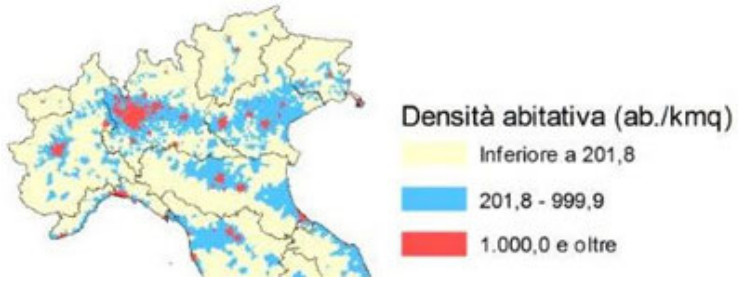

Figura 3.5: Densità abitativa fonte: Centro Documentazione e Studi Anci-Ifel su dati Istat, 2012

Parma, Piacenza, Modena, Reggio Emilia, Bologna, Forlì, Ferrara, Ravenna, Rimini e Cesena. Tali città corrispondono ai capoluoghi di provincia.

L'analisi spaziale permette di assegnare ogni Tweet a uno dei gruppi in base alla sua vicinanza ai diversi centri. Una volta suddivisi i Tweet si richiama il filtro quantitativo e si determina se il numero di Tweet rilevati appartenenti ad un cluster è un outlier; per stabilire se un valore è un outlier, si confronta il numero di Tweet pubblicati in un gruppo con il numero di Tweet pubblicati nello stesso gruppo, nello stesso intervallo di tempo, nei giorni precedenti; questo modulo, come il precedente, viene richiamato due volte con due intervalli di tempo differenti. Qualora il valore fosse elevato e fosse quindi un outlier, possiamo dire che sta accadendo qualcosa di eccezionale in quella zona.

3.5 Pagina web di controllo

Questo è il modulo più semplice e intuitivo. Esso è stato creato sotto forma di pagina web, la quale mostra visivamente i dati contenuti nel database e l'elaborazione delle analisi effettuate dal sistema di rilevamento. Per la parte di server-side si è utilizzato php, e si è fatto uso dell'estensione PDO[3] per interfacciare il server con il database MySql in cui sono contenuti i nostri

[3]L'estensione PHP Data Objects (PDO) definisce un'interfaccia leggera e coerente per l'accesso ai database in PHP. Sito web:http://php.net/manual/en/book.pdo.php

dati. La parte client, visivamente, è composta da 5 aree: una parte grafica, una mappa, il testo dei Tweet, la word cloud ad essi associata e il banner.

3.5.1 Grafico

La prima area che vediamo mostra un grafico che presenta in ascissa l'orario e in ordinata il numero di Tweet. Si osserva graficamente, dunque, l'andamento del numero dei Tweet pubblicati nel tempo, nel quarto d'ora. Per realizzarla si è fatto uso della libreria javascript Chart.js[4] che permette la creazione di grafici; Nella figura 3.6 è riportato un esempio dell'area appena

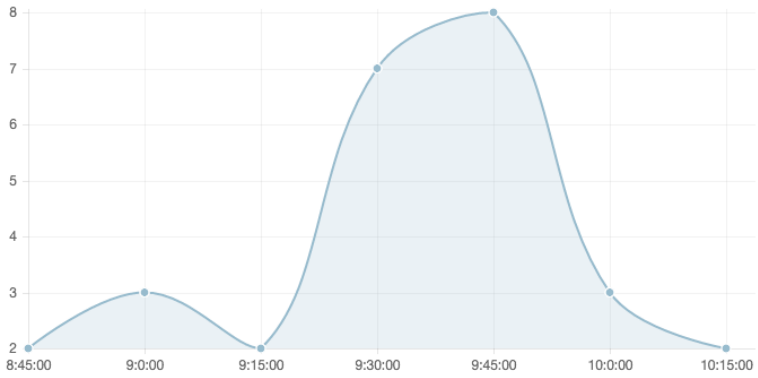

Figura 3.6: Grafico della Pagina web di Controllo

descritta e mostra l'andamento del numero dei Tweet dalle 8:45 alle 10:15, vi è un picco alle ore 9:45, ma è solo apparente in quanto, se si va a controllare l'andamento storico dei Tweet in quell'ora, spesso si ha un valore simile a quello mostrato in figura.

3.5.2 Mappa

In questa area viene mostrata la localizzazione dei Tweet rilevati nell'ultimo quarto d'ora. I Tweet, identificati da un indicatore di mappa, sono disposti in base alla loro geolocalizzazione, in una mappa. Tale mappa è centrata nella zona Nord-Est dell'Italia, siamo così in grado di osservare la

[4]sito web: http://www.chartjs.org/

regione di nostro interesse, l'Emilia Romagna. Si può aumentare o diminuire l'ingrandimento della cartina in modo da poter scegliere che sezione osservare della regione. La figura 3.7 mostra la mappa presente nella pagina di con-

Figura 3.7: Mappa della Pagina web di Controllo

trollo, in essa, facendo click sull'indicatore di mappa di interesse, è possibile visionare il Tweet relativo rappresentato nella mappa.

3.5.3 Tweet

L'area definita "Tweet" ha lo scopo di mostrare il testo e l'autore dei Tweet pubblicati nell'ultimo quarto d'ora. Tale sezione permette di leggere il contenuto dei Tweet che il nostro sistema seleziona, consentendoci di verificare, nel caso in cui venga registrato un picco, in che cosa consiste l'evento registrato. Nella figura 3.8 ne viene mostrato un esempio.

Figura 3.8: Elenco Tweet della Pagina web di Controllo

3.5.4 Word Cloud

La word cloud mostra le parole che compaiono con maggior frequenza nei Tweet. Per la sua realizzazione si è fatto uso di jQCloud[5], un plugin di jQuery, che permette, fornendogli una lista di parole e la loro ricorrenza, di creare uno schema word cloud. La figura 3.9 mostra un esempio di Word

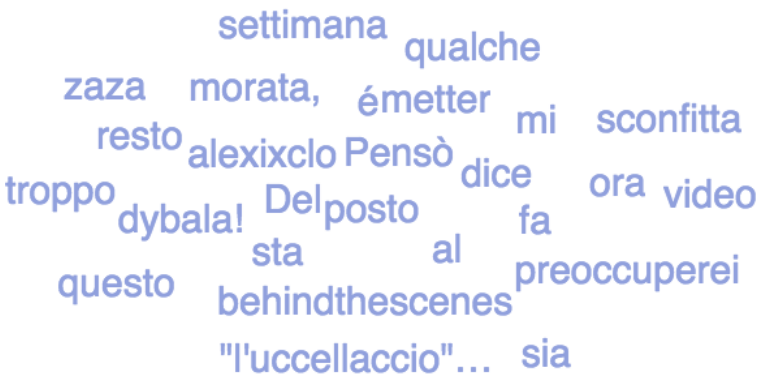

Figura 3.9: Word Cloud della Pagina web di Controllo

Cloud generata dai messaggi pubblicati dagli utenti nell'ultimo quarto d'ora;

[5]sito web: http://mistic100.github.io/jQCloud/index.html

come possiamo notare al momento in cui è stata presa l'immagine, non vi è alcuna parola che risalti più delle altre, questo è sintomo del fatto che non vi è un evento in corso.

3.5.5 Banner

In ultimo, nella parte superiore della pagina, vi è un banner il quale informa, in modo efficace, se il sistema sta funzionando correttamente e se è stato rilevato un evento.

L'immagine 3.10 mostra la pagina in esame.

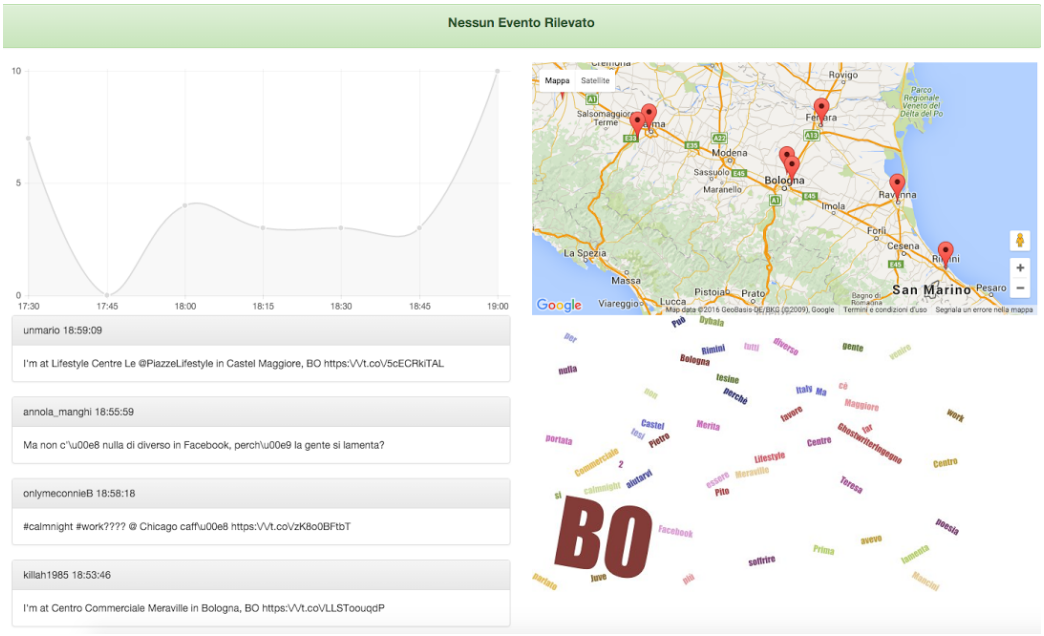

Figura 3.10: Pagina web di Controllo

Capitolo 4

Simulazioni

Durante il periodo di attività del nostro sistema di rilevamento (attivo dal 01/01/2016) nella regione Emilia Romagna, non sono accaduti eventi straordinari, questo ci ha portato a decidere di simulare l'accadere di un evento al fine di testare l'effettiva efficacia del nostro sistema. Il primo passo da fare è stato decidere quale numero di Tweet scaturisse dall'accadere di un evento, per questo abbiamo preso il dataset[22] nel quale sono presenti i Tweet relativi a 4 disastri naturali accaduti in Italia. Di seguito andiamo ad illustrare il metodo usato per la simulazione e i risultati da essa ottenuti.

4.1 Metodo

Il metodo che abbiamo ideato per la nostra simulazione riproduce l'accadere di un evento, nello specifico, il terremoto del Maggio 2012 che ha coinvolto l'Emilia Romagna, la Lombardia e il Veneto. Abbiamo usato un dataset composto da 5642 Tweet annotati manualmente in lingua italiana, questo dataset comprende i messaggi di Twitter riguardanti quattro differenti disastri naturali accaduti in Italia dal 2009 al 2014. La tabella 4.1 mostra la sua composizione.

I passaggi che abbiamo seguito per la simulazione sono i seguenti:

- Selezione dei Tweet dal dataset
- Inserimento dei Tweet nel nostro database

- Esecuzione dei moduli di analisi come se fossimo nel giorno dell'evento

Luogo	**Tipo**	**Anno**
Sardegna	*Inondazione*	2013
L'Aquila	*Terremoto*	2009
Emila	*Terremoto*	2012
Genova	*Inondazione*	2014

Tabella 4.1: Coposizione Dataset

Dal dataset abbiamo estrapolato i Tweet geolocalizzati in Emilia Romagna. Il passaggio successivo è stato inserire questi dati, in un giorno casuale, nel nostro database mantenendo l'ora in cui essi sono stati eseguiti. Abbiamo quindi lanciato i moduli di Analisi, dando come target il giorno scelto in precedenza.

Per vedere i tempi di risposta del nostro sistema, abbiamo deciso di traslare l'ora di inizio dell'evento da un minuto fino ad un massimo di 15. Abbiamo poi calcolato il tempo di risposta, minimo, medio e massimo del modulo di acquisizione. Tale valore lo andremo a sommare alle misure, fatte in precedenza, per simulare il tempo di risposta dell'intero sistema nei tre casi, questo è necessario perchè i primi dati non tenevano conto del modulo di acquisizione.

4.2 Risultati

I risultati ottenuti mostrano come il nostro sistema sia in grado di rilevare un evento. Abbiamo messo in relazione la differenza tra l'accadere dell'evento e la sua rilevazione da parte del sistema nei tre casi sopracitati (migliore, medio e peggiore) e possiamo vedere, figura 4.1, come nel caso migliore il nostro tempo di risposta medio sia stato di 506 secondi, nel caso medio di 519.70 mentre varia di molto nel caso peggiore in cui è stato di 605 secondi. Riguardo a questi dati bisogna tenere in considerazione che il nostro sistema

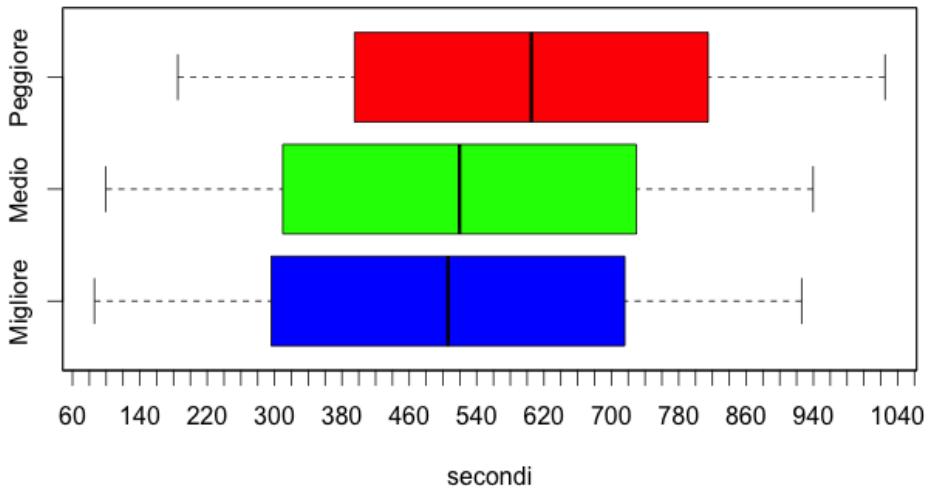

Figura 4.1: Box-Plot dei tempi di risposta

è eseguito dalla piattaforma hardware RaspberryPi modello A[1].

Gli eventi salienti del terremoto del 2012 possono essere riassunti in 3 passaggi dei quali mostriamo i relativi word cloud: ore 4:03 prima scossa figura 4.2, ore 4:23 prime impressioni riguardo al fatto figura 4.3, ore 5:03 scossa di assestamento figura 4.4 [25].

Figura 4.2: Word Cloud prima scossa

[1]https://www.raspberrypi.org/documentation/hardware/raspberrypi/

Figura 4.3: World Cloud prime impressioni

Figura 4.4: Word Cloud scossa di assestamento

Di questi passaggi il nostro sistema è stato in grado di seguire i primi due mentre non ha rilevato la seconda scossa. Siamo andati a guardare i Tweet del dataset per capirne il motivo e abbiamo notato come la maggior parte dei messaggi, riguardo all'assestamento, non siano geolocalizzati e quindi, da noi, non possano essere rilevati. Questa è una debolezza del nostro sistema che utilizza la geolocalizzazione come primo criterio di selezione.

Conclusioni

I Social Media sono lo strumento ormai preferenziale per esprimere le proprie opinioni, condividere fatti relativi alla propria vita ma anche per seguire gli eventi di attualità. Non solo gli enti governativi e le organizzazioni ufficiali utilizzano tali mezzi per comunicare ma anche i singoli cittadini, pubblicando ciò che sta loro accadendo, possono essere visti come sensori. Il Social Sensing, infatti, si basa sul presupposto che un gruppo di persone possa fornire un set di informazioni paragonabili a quelle fornite da un singolo sensore e hanno come scopo il poter rilevare tempestivamente eventi di preoccupazione sociale [2].
Nello studio della letteratura si sono visti molti esempi di applicazioni che sfruttano i Social Media per rilevare e migliorare la gestione di eventi catastrofici. Attraverso il monitoraggio di Twitter, selezionando il Tweet geolocalizzati in Emilia-Romagna, abbiamo dimostrato, tramite simulazione, come sia possibile rilevare l'accadere di un evento eccezionale. L'obiettivo di questo lavoro era mostrare come può essere semplice la creazione dello scheletro di una applicazione volta a monitorare l'accadere di eventi in una regione di interesse e dunque il risultato è stato raggiunto nonostante la necessità di migliorare l'implementazione del sistema creato.
Sicuramente alcuni aspetti da migliorare sono, il passaggio allo streaming dei Tweet per una maggior reattività del sistema perchè permetterebbe di ricevere le informazioni in tempo reale. Da questa miglioria dovrebbe poi conseguire un aumento della velocità di filtraggio perchè il client, impegnato in questa fase, non può ricevere altri Tweet. Per quanto riguarda il filtraggio, il miglioramento che si potrebbe apportare è una automatizzazione dei

filtri: ad esempio, blacklist degli utenti che si aggiorna autonomamente. Per quanto riguarda gli sviluppi futuri, dovrebbe essere implementato un modulo finale che vada ad operare a livello del testo dei Tweet, questo permetterebbe la classificazione delle informazioni in essi contenute.

Fra le difficoltà incontrate la maggiore è quella che riguarda i dati geolocalizzati, essi sono un numero molto esiguo, circa 300 Tweet ogni giorno per una regione di 4,451 milioni di abitanti, è un dato molto piccolo e rende difficoltosa l'identificazione di picchi.

Bibliografia

[1] Agarwal, A., Xie, B., Vovsha, I., Rambow, O., and Passonneau, R. "Sentiment analysis of twitter data." in: Proceedings of the ACL 2011 Workshop on Languages in Social Media, pp. 30-38, 2011.

[2] Marco Avvenuti, Stefano Cresci, Mariantonietta N. La Polla, Andrea Marchetti and Maurizio Tesconi, "Earthquake emergency management by Social Sensing" in: The Second IEEE International Workshop on Social and Community Intelligence, 2014

[3] Barbosa L., and Feng J. "Robust sentiment detection on twitter from biased and noisy data" in: Proceedings of COLING, pp. 36-44, 2010.

[4] Cameron, M. A., Power, R., Robinson, B., and Yin, J. "Emergency situation awareness from twitter for crisis management" in: Proceedings of the WWW 2012 Companion, ser. WWW 12 Companion. New York,NY, USA: ACM, pp. 695-698, 2012.

[5] Forman, G. "An extensive empirical study of feature selection metrics for text classification". The Journal of Machine Learning Research, 3, pp. 1289-1305, 2003.

[6] Go A., Bhyani R. and Huang L., "Twitter sentiment classification using distant supervision." CS224N Project Report, Stanford, 2009.

[7] Han Dong, Milton Halem, and Shujia Zhou, "Social media data Analytics Applied to Hurrucane Sandy" SocialCom/PASSAT/BigData/EconCom/BioMedCom 2013.

[8] Kaur Kumar "Sentiment analysis from social media in crisis situation" in: International Conference on Computing, Communication and Automation (ICCCA2015).

[9] Kouloumpis E., Wilson, T., and Moore, J. "Twitter sentiment analysis: The good the bad and the omg!" in: Proceedings of the ICWSM, 2011.

[10] Laurie BUTGEREIT: "CrowdSourced Weather Reports: An Implementation of the μ Model for Spotting Weather Information in Twitter " in: IST-Africa 2014 Conference Proceedings Paul Cunningham and Miriam Cunningham (Eds) IIMC International Information Management Corporation, 2014.

[11] Mandel B., Culotta, A., Boulahanis, J., Stark, D., and Lewis B. "A demographic analysis of online sentiment during hurricane Irene." in: NAACL-HLTWorkshop on Language in Social Media, 2012.

[12] Mejova,Y. "Sentiment Analisys: An overview", Comprehensive exam paper.

[13] Moschitti A. "Efficient convolution kernels for dependency and constituent syntactic trees." in: Proceedings of the European Conference on Machine Learning. pp. 318-329, 2006.

[14] Nagy, A., Stamberger, J.: "Crowd sentiment detection during disasters and crises." in: 9th International Conference on Information Systems for CrisisResponse and Management, ISCRAM, 2012.

[15] Pak, A., and Paroubek, P."Twitter as a corpus for sentiment analysis and opinion mining." in: Proceedings of LREC 2010, 2010.

[16] Pang B., Lee, L., and Vaithyanathan, S."Thumbs up: sentiment classification using machine learning techniques." in: Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10, Association for Computational Linguistics, pp. 79?86, 2002.

[17] Pennebaker, J., Booth, R., and Francis, M. "Liwc2007: Linguistic inquiry andword count, Computer software."Austin, TX: LIWC. Net, 2007.

[18] Tetsuhiro Sakai,Keiichi Tamura "Identifying Bursty Areas of Emergency Topics in Geotagged Tweets using Density-based Spatiotemporal Clustering Algorithm", 2014 IEEE 7th International Workshop on Computational Intelligence and Applications,November 7-8, 2014, Hiroshima, Japan.

[19] Tumasjan A., Sprenger, T., Sandner, P., and Welpe I. "Predicting elections with twitter: What 140 characters reveal about political sentiment.." in: Proceedings of the Fourth International AAAI Conference onWeblogs and Social Media (2010), pp. 178-185, 2010.

[20] Vasileios Lampos, Nello Cristianini "Tracking the flu pandemic by monitoring the Social Web" 2010 2nd International Workshop on Cognitive Information Processing

[21] Zheng Xu, Hui Zhang, Yunhuai Liu and Lin Mei: "Crowd Sensing of Urban Emergency Events based on Social Meida Big Data" in: IEEE 13th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom), 2014.

[22] S. Cresci, M. Tesconi, A. Cimino and F. Dell'Orletta. "A Linguistically-driven Approach to Cross-Event Damage Assessment of Natural Disasters from Social Media Messages" in: Proceedings of the 24th international conference companion on World Wide Web. ACM, 2015.

[23] "Twitter, ecco la storia di un successo", 07 Novembre 2013 http://www.corrierecomunicazioni.it/it-world/24153_twitter-ecco-la-storia-di-un-successo.htm

[24] Company, 31 Dicembre 2015 https://about.twitter.com/it/company

[25] "Terremoto Emilia 2012: i tweet che contano", http://www.datajournalism.it/terremoto-emilia-2012-i-tweet-che-contano/
Infografiche a cura di Maurizio Tesconi e Stefano Cresci, Istituto di Informatica e Telematica, Cnr di Pisa.