



BANDO INNOLABS

“Sostegno alla creazione di soluzioni innovative finalizzate a specifici problemi di rilevanza sociale”

Codice Progetto Y95B457



SBSoft



Deliverable B.3: Documento tecnico di progettazione del sistema di monitoraggio dei canali social

Data 26/02/2019

Versione 1.0

Deliverable B.3: Documento tecnico di progettazione del sistema di monitoraggio dei canali social

INFORMAZIONI GENERALI

Nome progetto	EasyPAL “Ecosistemi e Servizi Digitali in Cloud per i Cittadini e la PA Locale”
Ambito	BANDO INNOLABS “Sostegno alla creazione di soluzioni innovative finalizzate a specifici problemi di rilevanza sociale”
Riservatezza	Riservatezza ai sensi dell'art. 14 dell'atto di costituzione dell'ATS denominata “Easy PAL” registrata a Lecce in data 01/06/2018 al n. 5694/1T da Notaio Pellegrino.

RESPONSABILITA'

Funzione	Nome	Data
Redatto da	UNISALENTO – LABORATORIO DI INGEGNERIA ECONOMICO GESTIONALE	26/02/2019
Contributi di		
Controllato e approvato da		

STORIA DELLE VARIAZIONI

Revisione	Data emissione	Parte modificata	Descrizione della variazione
1.0	30/12/2018	Tutto il documento	Versione iniziale del documento
2.0	26/02/2019	Tutto il documento	Versione finale del documento

INDICE

1. Introduzione	4
2. Obiettivo del documento.....	5
3. Background teorico di riferimento	6
4. Definizione di un sistema di Sensori Sociali per il monitoraggio dei canali social	11
4.1 Definizione di uno Smart Social Sensor	11
4.2 Sorgenti di Big Data e tecniche di analisi per uno Smart Social Sensor	13
4.3 Esempi di Smart Social Sensor.....	15
5. Progettazione dell'architettura del sistema di monitoraggio	18
6. Data requirements per la realizzazione del prototipo di sistema di monitoraggio	22
Bibliografia	23

1. Introduzione

Uno degli effetti in cui le tecnologie digitali hanno trasformato il modo di comunicare, di lavorare e trascorre il tempo libero, è dato dal significativo aumento della mole di dati generati quotidianamente (Agarwal et al., 2014). Se da un lato, le quantità di dati sono ormai misurate in exabyte (10^{18}) e zettabyte (10^{21}), gli stessi si presentano con differenti nature e provenienze. Dai sensori usati per raccogliere dati sul clima ai rilevatori di traffico e alle informazioni sui voli, dai messaggi sui social media (es. Twitter e Facebook) alle immagini digitali e ai video, dalla registrazione delle transazioni economiche online fino ai segnali GPS dei telefoni cellulari, i dati generati da una molteplicità di piattaforme includono contenuto testuale (strutturato, semi-strutturato, non strutturato) e contenuti multimediali (testi, immagini, audio e video).

La crescente disponibilità di dati ha permesso alle organizzazioni di dotarsi di opportune metodologie e tecnologie in grado di elaborarli in tempo reale e direttamente all'interno del processo operativo o strategico nel quale i dati stessi sono presenti, al fine di generare input e considerazioni utili a migliorare i propri processi decisionali. Tutto questo rientra all'interno del paradigma dei Big Data e delle tecnologie di Data Analytics, le quali sono diventate una pratica largamente adottata allo scopo di elaborare grosse quantità di dati eterogenee per derivare informazioni preziose per supportare i processi decisionali (Sivarajah et al., 2017).

In questa prospettiva diventano rilevanti e attuali le applicazioni di *Mobile Crowd Sensing* che sfruttano la possibilità di rilevamento su larga scala, permettendo la partecipazione e la raccolta di dati implicita ed esplicita da dispositivi mobile e dai social network (Guo et al., 2014). Un'altra tipologia di applicazioni che sfruttano questi paradigmi emergenti data-intensive sono quelle di *Social Sensing* che si basano sull'idea che comunità o gruppi di persone possano fornire un insieme di informazioni simili a quelle ottenibili da un singolo sensore, generando in tempo reale una conoscenza complessa e adeguata di uno o più specifici problemi. Infine, una terza classe di applicazioni rilevanti nello scenario dei Big Data sono quelle di *Social Big Data* che integrano gli effort di ricerca sui domini dei social media con quelli in ambito Big Data al fine di fornire conoscenza sensibile e rilevante proveniente da fonti di dati di social media a qualsiasi utente.

Anche per le Pubbliche Amministrazioni Locali (PAL), la possibilità di ispirarsi a questi paradigmi utilizzando le tecnologie di Big Data Analytics attraverso il monitoraggio "in progress" di iniziative, servizi ed eventi, può contribuire alla creazione di nuova conoscenza. In Yoon et al. (2016) la dimensione di analisi in real-time consente un'ampia comprensione dei trend in atto. La

valutazione dei trend attraverso l'identificazione di discussion topics, opinion leader e il monitoraggio dei top influencer, l'estrazione del sentiment da blog, media social network e micro-blog possono portare la PAL ad acquisire in modo efficace la conoscenza delle opinioni e dei comportamenti dei cittadini, fino al punto in cui poter raccogliere la 'saggezza delle folle' e renderla operativa per la progettazione e l'erogazione di servizi centrati sul cittadino.

La capacità di monitorare tali fenomeni ha implicazioni dirette sulla possibilità di comprendere e descrivere gli eventi del mondo reale e le complessità delle dinamiche sociali, con applicazioni in svariati domini fra cui quello della pianificazione di smart-city (Quercia et al., 2011). Inoltre, l'analisi di eventi passati e presenti, anche monitorati attraverso i social media, potrebbe contribuire alla creazione di previsioni affidabili circa l'esito di un evento predeterminato o di eventi futuri non previsti. Studi recenti hanno dimostrato il successo delle previsioni sulla diffusione di malattie (Culotta, 2010), sull'aumento dei prezzi (Bollen et al., 2011), o sui risultati di competizioni politiche. Infine, poiché alcuni strumenti di comunicazione possono essere anche utilizzati per orientare il corso delle dinamiche umane online e offline, al fine di generare false notizie per supportare particolari opinioni, la rilevazione di attività anomale può aiutare a prevenire eventuali abusi sulle piattaforme social online.

2. Obiettivo del documento

Dato il significativo potenziale dell'impatto sociale previsto nel progetto dalla diffusione e l'utilizzo dei servizi della piattaforma EasyPAL, l'obiettivo del documento è quello di descrivere l'architettura di un sistema innovativo che, attraverso l'analisi di grandi quantità di dati provenienti da canali social, potrà alimentare e rendere accessibile un cruscotto di indicatori per misurare il capitale sociale che la piattaforma EasyPAL contribuirà a creare e consolidare nella PAL che lo adotta.

Più in particolare, in questo documento si descrive il concetto di *“sensore sociale”*, definito in questo contesto come agente software in grado di raccogliere in modo automatico e continuo i dati di interesse, elaborare analisi per estrarre in tempo reale informazioni a valore aggiunto. Il *“sensore sociale”* così definito, applicato al paradigma dei Big Data, ha portato alla progettazione di una piattaforma software che aggrega diversi sensori sociali e misura il capitale sociale sviluppato da una PA in termini di relazioni coi cittadini, aziende, altre PA e fornitori di servizi.

Questi indicatori saranno visualizzati in un cruscotto intelligente, sviluppato utilizzando tecniche di Big Data Analytics per l'implementazione di algoritmi di data mining per l'estrazione di informazioni allo scopo di fornire alla PA un set di servizi in grado di monitorare in tempo reale e in modo continuo differenti sorgenti di informazioni sociali.

La progettazione dell'architettura che realizza la piattaforma software e la selezione di tutte le sue componenti tecnologiche si basano sull'analisi delle migliori pratiche di sistemi di Big Data in quanto consentono di soddisfare i requisiti di capacità di analisi di grandi quantità di dati non strutturati e di calcolo e monitoraggio effettuati in tempo reale.

Dal punto di vista dei principi di design architetturale del sistema, si prevedono tre layer: il layer di *Data Management* per la raccolta dei dati; il layer di *Data Analytics* per l'implementazione dei meccanismi di estrazione di nuova conoscenza dai dati raccolti; il layer di *Data Visualization* per la visualizzazione degli insight prodotti dall'analisi in modo da renderli immediatamente fruibili all'utente.

3. Background teorico di riferimento

La crescita esponenziale di dati prodotti dalle organizzazioni e presenti sul web, generati da individui, sistemi informatici e dispositivi digitali, da un lato apre nuovi scenari e opportunità per la loro valorizzazione e crea al contempo nuove sfide tecnologiche e manageriali legate alla loro gestione, analisi, elaborazione ed interpretazione. Il paradigma dei Big Data è quindi emerso negli ultimi anni come possibile alternativa per acquisire, memorizzare, gestire e analizzare in modo efficace, flessibile ed efficiente i dati in contesti applicativi data-intensive, per offrire servizi innovativi in domini come quello sanitario, bancario, turistico, della formazione, della pubblica amministrazione e del manifatturiero intelligente.

Inoltre, la diffusione rapida dei Big Data dimostra quanto sia grande l'interesse che le organizzazioni hanno nei confronti di tale fenomeno e nella generazione di valore attraverso lo sviluppo e l'adozione di soluzioni basate su questo paradigma tecnologico.

In realtà, la gestione e l'analisi di grandi quantità di dati, raccolti in una forma strutturata e non strutturata, ed elaborati da algoritmi e tecniche di analisi innovative, possono aprire nuove sfide tecnologiche e organizzative che rendono le aziende e le organizzazioni più flessibili, aperte, creative, resilienti e competitive. Oggi, il paradigma dei Big Data sta rappresentando una nuova

ondata di innovazione tecnologica, opportunità commerciali e di sviluppo di competitività, generando un grande impatto in diversi campi quali sanità, e-commerce, ricerca scientifica e manifatturiero, finanza e istruzione, politica, servizi pubblici, e-government e sicurezza pubblica (Elia et al., 2018).

Esistono diversi modelli e definizioni di Big Data in letteratura. Laney (2001) ha formalizzato, per la prima volta, un modello descrittivo del paradigma dei Big Data conosciuto come il modello delle 3V, che più recentemente è stato arricchito da altri studiosi (McAfee & Brynjolfsson, 2012; Fosso Wamba et al., 2015). Mentre le classiche caratteristiche a 3V (Volume-Velocità-Varietà) presentano sfide significative per l'analisi dei dati, le ulteriori 2V (Veracità e Valore) sono quelle che offrono il potenziale di valore per le imprese di elaborare migliori decisioni di business e ottenere vantaggi competitivi. L'introduzione di queste due ulteriori dimensioni ha portato alla definizione del cosiddetto modello Big Data 5Vs. Quindi, il modello iniziale delle "3V" (Volume, Varietà, Velocità) si è evoluto nel tempo nel modello delle "5V", che aggiunge le dimensioni di Valore e Veracità come due ulteriori caratteristiche fondamentali dei Big Data.

Segue una descrizione più dettagliata di ciascuna componente del modello.

1. Il "Volume" (*data-at-rest*) si riferisce all'enorme mole di dati a disposizione, al vantaggio ottenuto dalla capacità di elaborare grandi quantità di informazioni, nonché alla tendenza di molte aziende di archiviare un'enorme quantità di diversi tipi di dati approvvigionandosi da fonti eterogenee (es. social network, dati sanitari, finanziari, biochimici e genetici, dati meteo).
2. La "Varietà" (*data-in-many-forms*) si riferisce al fatto che tali dati non hanno una struttura fissa e raramente si presentano in una forma perfettamente ordinata tale da essere immediatamente pronti per l'elaborazione. Infatti, i Big Data possono essere altamente strutturati se provenienti da database relazionali (es. elenco clienti, catalogo prodotti), o semi-strutturati (es. web log, feed di social media, dati grezzi di un sensore, e-mail, pagine HTML, documenti XML), o totalmente non strutturati (es. file multimediali come video, immagini, e tracce audio, dati provenienti dai social media, documenti testuali).
3. La "Velocità" (*data-in-motion*) si riferisce ai flussi e alla rapidità dei dati in ingresso e in uscita, e conseguentemente alla velocità per la loro elaborazione.

4. Il “Valore” (*data-in-evidence*) si riferisce all’obiettivo per il quale le soluzioni Big Data sono progettate, cioè creare ed estrarre valore da grandi volumi di dati, mediante un accesso diretto, interattivo e in tempo reale ad analisi e grafici che evidenziano nuove opportunità di business, che supportano il processo decisionale, o che forniscono input utili per progettare nuovi prodotti o servizi, o creare nuovi modelli di business.
5. La “Veracità” (*data-in-doubt*) si riferisce ai concetti di accuratezza, certezza, e precisione del dato. In ambito Big Data, la non veracità del dato può essere causata da incongruenze, approssimazioni, ambiguità, duplicazioni, incompletezze, spam, e latenza. Su grosse moli di dati, tali imperfezioni tendono a non compromettere i risultati delle analisi, anche se è conveniente assegnare una probabilità piuttosto che considerarli certi.

Recentemente il modello delle “5V” è stato ulteriormente esteso, includendo altre dimensioni e caratteristiche legate ai dati che risultano determinanti in base alle esigenze o al dominio di applicazione. Fra le dimensioni più utilizzate, si cita:

- la “Variabilità” (*data-in-change*), spesso accoppiata alla varietà, che sottolinea la variabilità di semantica del significato associato al dato, in base al linguaggio o ai protocolli di comunicazione utilizzati;
- la “Volatilità” (*data-relevancy*), o decadimento, riferita all’intervallo di tempo in cui il dato può essere considerato valido e rilevante, elemento importante soprattutto nel caso di analisi real-time;
- la “Visualizzazione” (*data-in-view*), che riguarda le modalità di rappresentazione efficace, attraverso l'utilizzo di diversi strumenti e formati visivi, delle informazioni e della conoscenza destinate ai diversi target di riferimento;
- la “Complessità” (*data-complexity*), riferita al numero di sorgenti dati interessate dall’analisi che determina le operazioni di raccolta, pulizia, memorizzazione ed elaborazione dei dati.

Nel 2016, a valle di un’analisi di letteratura sulle principali definizioni esistenti sui Big Data e sui principali temi di ricerca ad essi associati, i Big Data sono definiti come *“il patrimonio informativo caratterizzato da un elevato volume, velocità e varietà, che richiede specifiche tecnologie e metodi di analisi per la loro trasformazione in valore”* (De Mauro et al., 2016). Tale definizione evidenzia le tre dimensioni chiave associate ai Big Data, cioè:

- "Volume", "Velocità" e "Varietà", che descrivono le caratteristiche principali dei dati;
- "Tecnologie" e "Metodi analitici", che descrivono gli strumenti necessari per utilizzare correttamente tali informazioni;
- "Valore", che descrive la trasformazione dei dati di partenza in informazioni e suggerimenti utili che possono creare valore economico e sociale per le aziende, le organizzazioni, e le comunità.

Infine, all'interno del paradigma Big Data, vi è un ulteriore e rilevante aspetto relativo ai dati da tenere in forte considerazione, e che assumerà sempre più rilievo nel prossimo futuro: la *privacy*.

Nonostante essa non sia ancora entrata pienamente nei modelli di definizione e caratterizzazione dei Big Data, la gestione della riservatezza dei dati può influenzare la fase di creazione dei dati (pezzi di informazioni nascoste), di analisi dei dati (aggregazioni o correlazioni di dati possibili solo accedendo a dati privati), e di eliminazione dei dati (cancellazioni di dati individuali che potrebbero generare incoerenze a livello dei dati aggregati).

Dalle definizioni e dai modelli sopra descritti e dalle dimensioni analizzate, risulta evidente la complessità del paradigma dei Big Data ed al tempo stesso l'enorme potenziale di business ad esso associato. Rispetto ai dataset tradizionali, nei Big Data l'informazione utile è nascosta dalla quantità di dati, rilanciando la sfida di riuscire ad individuare ed estrarre quello che può realmente creare valore. Recentemente, le attività di ricerca sono state focalizzate sia su questioni tecnologiche che su una dimensione orientata al "valore", che rappresenta la reale aspettativa e l'impatto sul business dell'intero paradigma dei Big Data che comprende aspetti tecnologici, economici ed organizzativi.

Una particolare categoria di Big Data che rappresenta circa il 90% dei Big Data esistenti (Gandomi et al., 2015) è rappresentata dai Big Social Data (BSD). Con questo termine si indicano grandi volumi di dati generati dagli utenti e che descrivono il comportamento e le interazioni sociali, in ambienti digitali (Olshannikova et al., 2017). I Big Social Data sono quindi dati generati sui social media e caratterizzati da alto volume, alta velocità, alta varietà, alto valore e alta variabilità. Dall'analisi della letteratura emerge come i BSD siano stati definiti come:

- scienza che studia le relazioni tra dati fisici del mondo reale e dati sui social media (Ishikawa, 2015);

- scienza sociale che, attraverso metodi di analisi quantitativa, studia le connessioni tra i flussi di informazioni e il comportamento umano (Pentland, 2014);
- risultato di approcci basati sul social computing (Guellil, 2015; Nguyen, 2015; Tang, 2014);
- risultato di ciò che deriva dalle pratiche comunicative mediate dalla tecnologia nella vita quotidiana (Coté, 2014);
- approccio orientato alla conoscenza della società e allo studio di discipline umanistiche (Burgess & Bruns, 2012) per migliorare la governance e l'organizzazione della società (Housley et al., 2014);
- metodologia basata sulla combinazione di tre elementi quali (Bello-Organ et al., 2016) i Big Data come paradigma di elaborazione, i Social Media come principale fonte di dati, e la Data Analysis come metodo per acquisire nuove conoscenze per utenti o aziende.

Emerge inoltre quanto il concetto di Big Social Data sia stato esplorato da diverse prospettive (ad esempio i social media, i social network online, il social computing e le scienze sociali computazionali) e come la ricerca abbia prestato poca attenzione a chiarire il concetto stesso di BSD e a comprenderne i fenomeni correlati preferendo, invece, concentrarsi sugli usi possibili dei BSD, ossia sull'estrazione di informazioni dai dati provenienti dai social media e dalle interazioni sociali online degli utenti a fini descrittivi o predittivi per gestire il processo decisionale umano in vari domini applicativi.

Esistono, quindi, differenti interpretazioni e termini con cui in letteratura ci si fa riferimento all'aspetto "sociale" nei Big Data. I termini più diffusi sono "Social Big Data" (SBD) e "Big Social Data" (BSD), che spesso sono utilizzati in maniera intercambiabile e comunque risultano non chiaramente ed unicamente definiti.

Fra le definizioni presenti in letteratura quella scelta come riferimento in questo contesto, è quella data da Olshannikova et al. (2017), che descrive i BSD caratterizzandone proprietà, modalità di generazione e possibili usi: *"Big Social Data is any high-volume, high-velocity, high-variety and/or highly semantic data that is generated from technology-mediated social interactions and actions in digital realm, and which can be collected and analyzed to model social interactions and behavior"*.

4. Definizione di un sistema di Sensori Sociali per il monitoraggio dei canali social

Questo paragrafo ha l'obiettivo di definire uno strumento a servizio delle PAL per il monitoraggio "in progress" di iniziative, servizi ed eventi, insieme alla relativa comprensione dei trend in atto e della conoscenza delle opinioni e dei comportamenti dei cittadini. A tal fine, si procederà con la definizione del concetto di *"Smart Social Sensor"* (Sensore Sociale Intelligente), identificando le sorgenti dati da considerare come input del processo di estrazione di conoscenza, nonché le tecniche di analisi e di data visualization utilizzate dal Sensore Sociale per la produzione e la rappresentazione di nuove informazioni e conoscenza.

4.1 Definizione di uno Smart Social Sensor

I sensori fisici in un ambiente reale producono un'enorme quantità di dati ambientali. Essi misurano e producono dati fisici o chimici, come misure di luce, calore, temperatura e umidità.

Quello del Sensore Sociale è un concetto derivato da quello dei sensori fisici, applicato nel contesto dei Social Media. Un Sensore Sociale può misurare e produrre, ad esempio, contenuti testuali, informazioni su utenti, dati geo-referenziati relative a posizioni. Un sensore sociale è quindi progettato per misurare dati e per generare informazioni in real-time per uno specifico compito analitico (Vasudevan et al., 2013; Matsuo et al., 2010).

Un Sensore Sociale nel contesto del progetto EasyPal è definito come un agente *"Smart"* capace di estrarre e gestire dati, elaborare informazioni e produrre nuova conoscenza a partire da una sorgente di social media data, utilizzando il paradigma computazionale dei Big Data.

Uno *"Smart Social Sensor"* è quindi il risultato di:

- approcci di *"Social Sensing"* basati sull'idea che comunità o gruppi di persone possono fornire set di informazioni simili a quelle ottenibili da un singolo sensore fisico, al fine di generare in real-time conoscenza complessa e adeguata alla risoluzione di uno o più problemi specifici;
- tecniche di *"Social Big Data"* basate sull'analisi di grandi quantità di dati provenienti da molteplici sorgenti distribuite, ma con un focus prevalente sui social media, al fine di fornire conoscenza rilevante per gli utenti.

Dal punto di vista logico-funzionale, uno Smart Social Sensor è costituito da tre componenti logico-funzionali, come rappresentato in Figura 1.

La componente di *Data Management* gestisce la fase di acquisizione ad alta velocità dei dati da una varietà di fonti (web, database relazionali, database NoSQL, sorgenti social, etc.), implementando opportunamente i diversi protocolli per l'accesso ai dati. Questa componente costituisce il filtro attraverso cui è possibile memorizzare solo i dati ritenuti utili o i dati "grezzi" con un minor grado di incertezza. Per alcuni Smart Social Sensor, in contesti applicativi particolari, la componente di Data Management gestisce anche la memorizzazione dei metadati che sono associati ai dati, e che quindi possono rivelarsi utili per ulteriori analisi. Infine, tale componente gestisce anche i processi di "pulizia" dei dati attraverso l'implementazione di un processo di ETL (Extraction, Transformation, Loading), al fine di renderli funzionali ad una loro elaborazione (strutturata o semi-strutturata), integrazione e conservazione nella struttura dati più appropriata e idonea (es. data warehouse, data mart, database operazionali, database NoSQL).

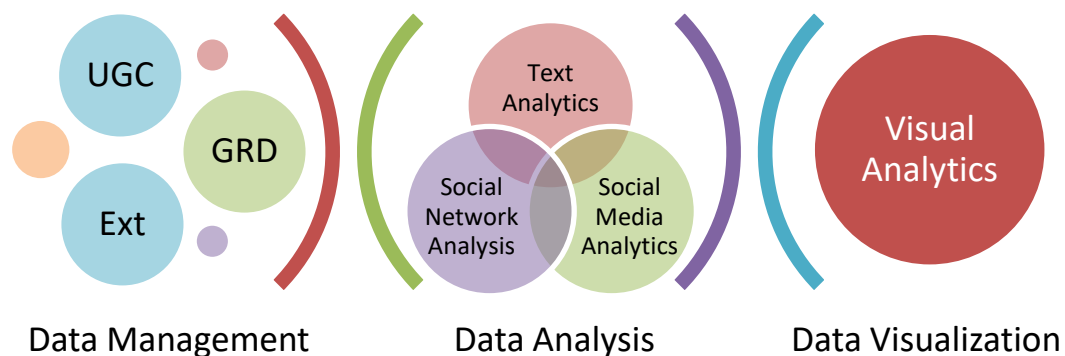


Figura 1. Schema logico-funzionale di uno Smart Social Sensor

La componente di *Data Analysis* realizza per ogni Smart Social Sensor l'esecuzione di query, la costruzione di modelli decisionali, e l'implementazione di algoritmi per identificare ed estrarre nuove informazioni e indicazioni. L'utilizzo di tecniche di data mining su data set integrati, puliti e affidabili permettono di migliorare la qualità e l'attendibilità dei dati, comprendere la semantica, e fornire funzioni di interrogazione intelligenti.

La componente di *Data Visualization* ha la responsabilità in uno Smart Social Sensor dell'implementazione del paradigma della Visual Analytics, che consiste nel combinare analisi automatiche con modalità di visualizzazione interattive per abilitare una comprensione effettiva dei dati, utilizzare modalità di reasoning riproducibili, e supportare così il processo decisionale.

Prendere decisioni efficaci che generano valore è l'obiettivo ultimo delle tecniche di analisi dei Big Data. Il raggiungimento di questo obiettivo passa per l'impiego e l'utilizzo di altrettanto efficaci tecniche di visualizzazione dei risultati generati dalle analisi stesse.

Uno Smart Social Sensor è sempre “special-purpose”, poiché progettato e realizzato per uno specifico obiettivo in termini di generazione di conoscenza, ed è caratterizzato da cinque elementi:

- una sorgente dati,
- un set di processi di data management,
- un set di processi di data analysis,
- un set di metafore di visualizzazione,
- un template di parametri di configurazione di tali processi.

4.2 Sorgenti di Big Data e tecniche di analisi per uno Smart Social Sensor

Uno Smart Social Sensor è progettato per assolvere ad un complesso e preciso task di produzione di nuova conoscenza che comprende diversi step in termini di estrazione di dati, elaborazione e visualizzazione dei risultati.

Pertanto, a partire dall'output desiderato, la progettazione e la successiva prototipazione di uno Smart Social Sensor terrà conto della tipologia di sorgenti dati da analizzare, del formato e della struttura dei dati in input per il processo di estrazione di informazioni, della scelta degli algoritmi per il processo di analisi, nonché della metafora di visualizzazione dei risultati.

La tipologia di Big Data considerata per il progetto EasyPAL sono i Big Social Data che Olshannikova et al. (2017) descrivono, caratterizzandone proprietà, modalità di generazione e possibili usi, come *“quei dati caratterizzati da alto volume, alta velocità, alta varietà e/o alto grado di semantica generati da interazioni e azioni sociali mediate dalla tecnologia in ambito digitale e che possono essere raccolti e analizzati per modellare comportamenti ed interazioni sociali”*.

Le sorgenti di Big Social Data per uno Smart Social Sensor considerate nell'ambito del progetto EasyPAL sono di due tipologie: i Social Networking Sites (SNS) e blogs, e i Corporate websites (piattaforme/applicazioni per la diffusione delle informazioni, per effettuare transazioni economiche, e per la condivisione di opinioni). I SNS sono luoghi online specificamente indirizzati a incoraggiare gli utenti a esprimere le proprie opinioni su qualsiasi tipo di argomento. Pertanto, le informazioni che contengono sono in qualche modo un riflesso di ciò che accade nella società. Per

questo motivo, viene prestata maggiore attenzione ai SNS come fonti di dati massivi potenzialmente utili nella previsione di variabili sociali.

A titolo di esempio, i Social Networking Sites più utilizzati come sorgenti di Big Social Data sono Twitter e Facebook.

Twitter è uno dei più popolari, con 332 milioni di utenti attivi mensilmente e che inviano in media oltre 500 milioni di tweet al giorno. Questa enorme quantità di informazioni "generate dall'utente", è stata utilizzata per prevedere eventi sociali ed economici sia presenti che futuri. L'analisi del contenuto dei tweet, infatti, ha contribuito a descrivere le preferenze politiche e ha previsto i risultati delle elezioni, ha previsto i movimenti del mercato azionario, ha monitorato l'opinione pubblica sulle nuove politiche.

L'altro Social Networking Sites più popolare ed utilizzato è Facebook che, con circa 1.650 milioni di utenti attivi, rappresenta senza dubbio un'altra fonte importante di dati per l'analisi dei comportamenti sociali ed economici. Tuttavia, tali dati sono anche più difficili da recuperare e analizzare a causa della loro eterogeneità e possibilità di intervento dell'utente in termini di visibilità e privacy. Nonostante ciò, numerosi studi hanno dimostrato come sia possibile utilizzare i contenuti generati dagli utenti di Facebook per determinare i profili di consumo, utili per scopi di marketing, prevedere i risultati elettorali e l'orientamento politico della popolazione.

Il formato dei contenuti in input ad uno Smart Social Sensor può essere:

- Strutturato: dati semplici da identificare ed estrarre, interrogare e memorizzare (catalogo prodotti, elenco clienti, etc.);
- Semi-strutturato: dati non organizzati in modelli predefiniti come database relazionali, che richiedono quindi regole specifiche per la loro cattura (pagine HTML, documenti XML, etc.);
- Non strutturato: dati che non possiedono uno specifico formato (file multimediali, dati provenienti dai social media, documenti di testo, etc.).

Uno Smart Social Sensor può avvalersi dell'uso di differenti tipologie di data store per la memorizzazione dei dati, e può implementare processi per eliminare dati incompleti o errati, per trasformare i dati di input in una forma più adatta all'analisi, per caricare e normalizzare in strutture relazionali i dati elaborati minimizzando la ridondanza. Inoltre, i processi di elaborazione dei dati per le successive fasi di analisi possono essere in modalità batch o real-time.

In funzione dei risultati dell'analisi sulle sorgenti dati, uno Smart Social Sensor può utilizzare un set di algoritmi per l'analisi dati, e modellare così il processo di estrazione delle informazioni.

Tra le varie famiglie di tecniche di analisi dati, quelle in elenco sono quelle che possono essere incorporate in uno Smart Social Sensor per poter essere utilizzate:

- Opinion Mining: usate per l'analisi di opinioni, "sentiment", giudizi ed emozioni di utenti verso entità tangibili o intangibili;
- Topic Modeling: utilizzate per l'identificazione e l'estrazione di temi o topiche da grandi corpora testuali;
- Engagement Analysis: usate per la valutazione del coinvolgimento di un utente verso un servizio, un prodotto, un evento o un'attività (ad es. quanto un consumatore nutre interesse verso un brand, oppure un cittadino verso la vita politica);
- Predictive Analysis: usate per l'analisi di dati e di informazioni storiche per comprendere il possibile comportamento di un'entità (es. individuo, gruppo, comunità, eventi);
- Social Network Analysis: usate per la modellazione di connessioni ed attività, nonché della dinamica di crescita, di sviluppo e di comportamento di singoli individui e di gruppi all'interno delle piattaforme di social media;
- Trend Analysis: utilizzate per scoprire i trend di settore, come le richieste dei prodotti, le intuizioni dei consumatori e la qualità del servizio di un settore.

Infine, ogni sensore sociale implementa la metafora di visualizzazione più adatta, al fine di facilitare il processo decisionale. Al fine di raggiungere questo obiettivo è richiesta infatti una buona visualizzazione dei risultati delle analisi stesse attraverso metodi e strumenti di supporto per creare immagini, diagrammi o animazioni utili a comunicare, comprendere e migliorare i risultati delle analisi sui Big Data. Le tecniche di visualizzazione dei dati che saranno prese in considerazione riguardano le Tag Cloud, i Clustergram, i flussi storici e i flussi spaziali di informazioni.

4.3 Esempi di Smart Social Sensor

In questo paragrafo sono presentati due esempi di Smart Social Sensor. Il primo esempio è relativo al monitoraggio continuo di contenuti testuali generati su un Social Network per l'estrazione in tempo reale del "sentiment", o più in generale delle opinioni, espresso dagli utenti all'interno dei

contenuti testuali da essi prodotti. Le tecniche di data analysis per la determinazione del sentiment fanno riferimento ad algoritmi supervisionati di classificazione multi-classe.

Lo Smart Social Sensor rappresentato in Figura 2 ha l'obiettivo di classificare i messaggi (tweet) in base alla polarità (positivo, neutro, negativo) del sentiment verso un argomento specifico (topic).

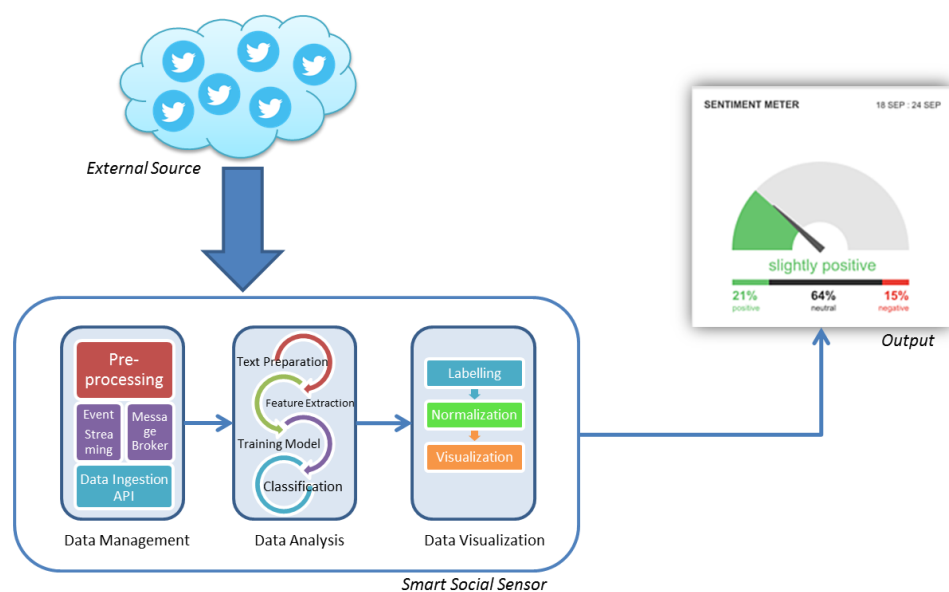


Figura 2. Schema di uno Smart Social Sensor per l'estrazione del sentiment da Twitter

Per quanto riguarda i processi di *Data Management*, la sorgente di Big Data per questo Smart Social Sensor è il Social Network Twitter che fornisce contenuti testuali, chiamati tweet, e delle API (Application Programming Interface) per il recupero dei contenuti. Il "parameter template" previsto per lo Smart Social Sensor, indica la Twitter Streaming API quale interfaccia a cui collegarsi per il recupero dei dati. Stabilire una connessione alle API di streaming di Twitter significa effettuare una richiesta HTTP di lunga durata e analizzare la risposta in modo incrementale. Nella richiesta, si possono filtrare i tweet per parola chiave, utente e area geografica specifica, lasciando aperta la connessione HTTP e raccogliendo quanti più tweet possibile. Il Message Broker è responsabile della gestione dello streaming dei dati sullo Smart Social Sensor, catturando come "evento" qualsiasi nuovo dato prodotto dal Social Network e memorizzando in una coda asincrona gli eventi ricevuti. Il componente di pre-processing dei dati applica, se necessario, una serie di tecniche per convertire i dati grezzi in un set di dati puliti e pronti per l'analisi (es. rescaling, normalizzazioni, rimozione di rumore, conversioni di numeri in testo, etc.).

I processi di *Data Analysis* di questo Smart Social Sensor, processano i testi recuperati dalla sorgente dati in modo da poterne estrarre un vettore di “feature” che lo rappresentano attraverso l’applicazione di tecniche quali la tokenizzazione, la lemmatizzazione, e la rimozione di stop words. I tweet così preparati vengono trasformati in vettori di termini (features) e sono pronti per la classificazione rispetto a un training set di testi già classificati in base al sentiment. L’output è quindi una probabilità che il testo analizzato appartenga a una delle classi (positivo, negativo, neutro) oggetto di classificazione.

I processi di *Data Visualization* raccolgono l’output del processo di Data Analytics per la visualizzazione dei risultati in real-time applicando la metafora di visualizzazione opportuna, come ad esempio i dashboard, i pie-chart o gli istogrammi.

Il secondo esempio è relativo ad uno Smart Social Sensor che identifica i “trend topics” da una pagina pubblica del social network Facebook. In Figura 3 è rappresentato il funzionamento di questo sensore, che esamina i testi contenuti nei post pubblicati dagli utenti.

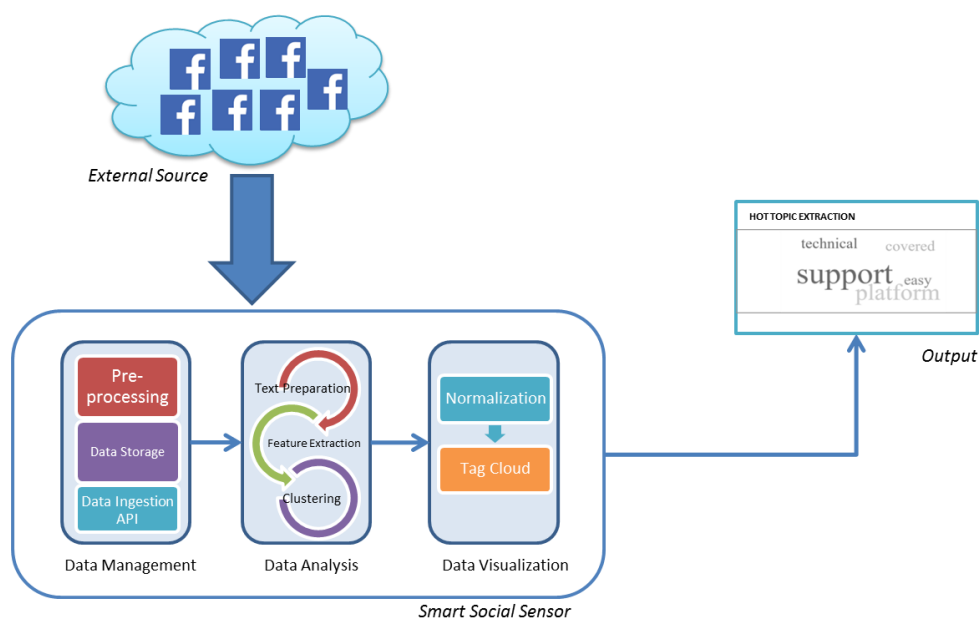


Figura 3. Smart Social Sensor per l'estrazione di trend topic da Facebook

Un “trend topic” è l’argomento di tendenza nelle discussioni all’interno di comunità o di gruppi di utenti. Identificare un trend topic con precisione permette di capire ciò che la gente pensa sia degno di nota o importante, e ciò che le persone trasmettono e condividono. Le organizzazioni che acquisiscono tale conoscenza possono prendere decisioni in tempo reale molto efficaci perché in linea con ciò di cui gli utenti dibattono e condividono.

Per quanto riguarda i processi di *Data Management*, il sensore rappresentato in Figura 3 estrae i contenuti dalla pagina del Social Network, memorizzandone contenuto e metadati. Il “parameter template” previsto per questo Smart Social Sensor, descriverà le API da usare per l'estrazione dei contenuti, la frequenza di interrogazione delle API, la pagina target. Come nel caso precedente, il componente di pre-processing dei dati applica, se richiesto, una serie di tecniche per convertire i dati grezzi in un set di dati puliti e pronti per l'analisi.

I processi di *Data Analysis* di questo Smart Social Sensor, processano i testi recuperati dalla sorgente dati in modo da poterne estrarre un vettore di “feature” che lo rappresentano. Sulla base delle feature estratte, un algoritmo di Clustering calcola i cluster di termini che rappresentano i trend topic. Il numero di cluster e le modalità di computazione sono descritte nel parameter template del sensore.

Le informazioni così ricavate saranno utilizzate dai componenti di *Data Visualization* per la loro rappresentazione. Per questo sensore, la metafora di rappresentazione prescelta è quella della tag cloud.

5. Progettazione dell'architettura del sistema di monitoraggio

In questo paragrafo, è descritto il design dell'architettura della piattaforma software basata su paradigma dei Big Data e dei micro servizi per la realizzazione dei Sensori Sociali.

L'architettura si ispira al paradigma delle architetture a micro-servizi, al fine di garantire resilienza e scalabilità dell'intero sistema. Infatti i principi alla base del paradigma sono:

- 1) un servizio per ogni elemento funzionale: esposizione di interfacce RESTful;
- 2) scalabilità del micro-servizio, replicandolo sull'infrastruttura in funzione della reale necessità;
- 3) gestione del dato decentralizzata;
- 4) gestione dell'infrastruttura decentralizzata.

La Figura 4 illustra l'architettura a micro-servizi per il sistema di monitoraggio dei canali social da realizzare nell'ambito del progetto.

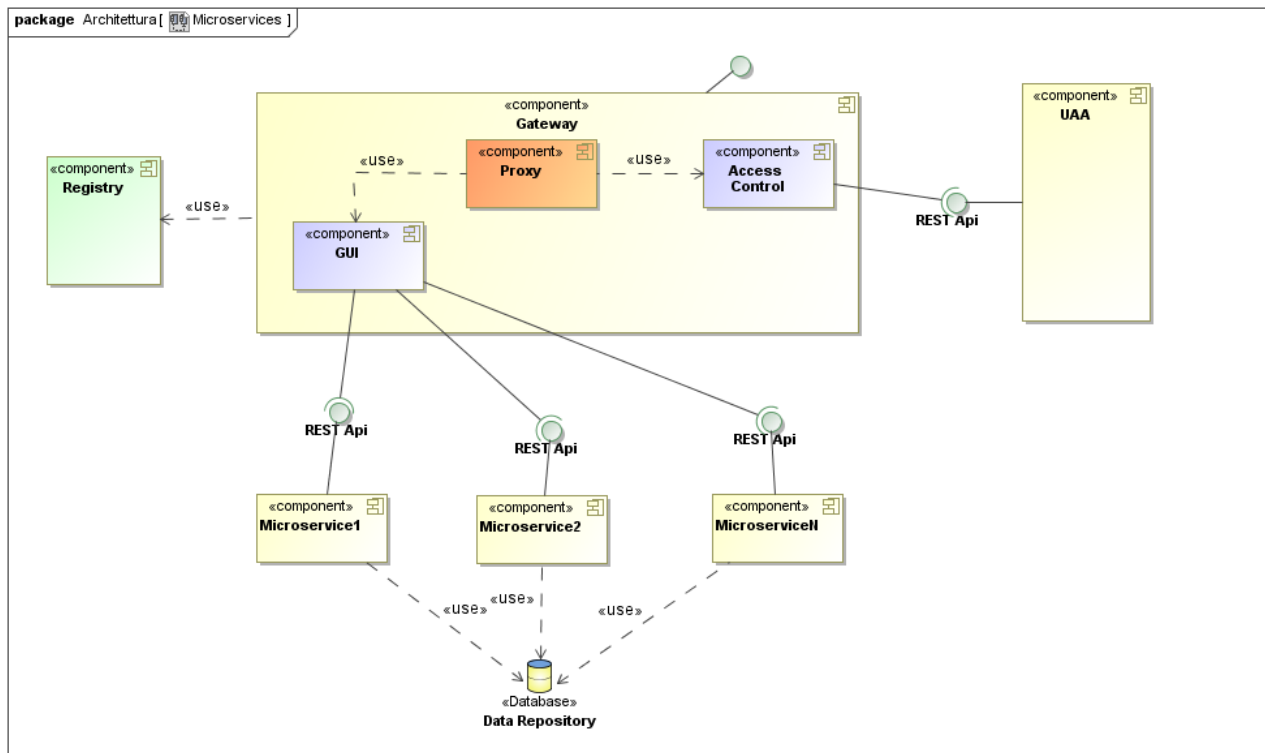


Figura 4. Architettura per il sistema di monitoraggio

Il componente di Registry permette ai componenti dell'architettura di registrarsi e leggere a runtime le proprie configurazioni. Inoltre, fornisce monitoraggio a runtime dello stato dei componenti stessi.

Il Gateway è il punto di ingresso per gli utenti all'applicazione, ed ha la responsabilità di:

1. smistare le richieste sul micro-servizio che realizza le funzionalità desiderate, attraverso il componente di proxy;
2. realizzare l'interfaccia utente attraverso il componente di GUI;
3. interfacciarsi con il componente esterno di User Authorization and Autentication.

Quest'ultima funzionalità è realizzata dal componente di UAA che gestisce l'identità degli utenti e il controllo degli accessi.

I micro-servizi realizzano le singole unità funzionali, sono "stateless" e permettono di eseguire diverse proprie istanze in parallelo.

La Figura 5 mostra la situazione di un numero diverso di istanze di micro-servizi eseguite in relazione al diverso carico che ogni servizio deve gestire. La differenza sostanziale rispetto ad

applicazioni con architetture “monolitiche” è che queste ultime hanno necessità di ridondare l’intera applicazione al fine di scalare, indipendentemente dalle funzionalità con maggiore carico.

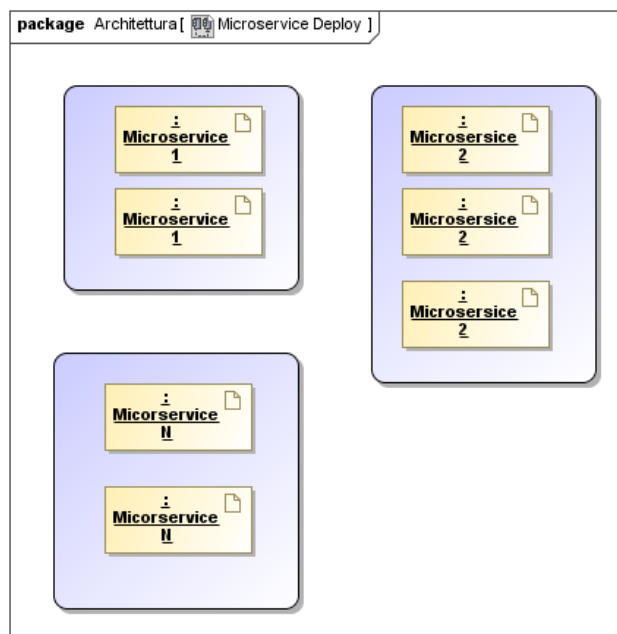


Figura 5. Architettura a micro-servizi della piattaforma

Quindi, i micro-servizi realizzano la logica di back-end delle funzionalità messe a disposizione dalla piattaforma di monitoraggio, e la GUI all’interno del Gateway realizza la dashboard per gli indicatori messi a disposizione dagli Smart Social Sensor. Ogni Smart Social Sensor è costituito da tre layer che realizzeranno le funzionalità di Data Management, Data Analysis e Data Visualization.

In particolare, i layer di Data Management e Data Analysis sono implementati da un unico micro-servizio.

Il layer di Data Management implementa, attraverso gli opportuni strumenti e tecnologie, i processi di ingestione, memorizzazione e preparazione dei dati per i sensori sociali.

Il layer di Data Analysis prevede la definizione dei motori di analisi dati per i sensori sociali, le interazioni con i repository di modelli predittivi utilizzati in fase di estrazione della conoscenza e con i Data Repository su cui saranno memorizzati i risultati delle analisi. Ogni motore di analisi è progettato in relazione alla funzione svolta e alla tipologia di Sensore Sociale al fine di contribuire in modo dettagliato e circostanziato all’estrazione delle informazioni utili alla costruzioni del set di indicatori. Ogni micro-servizio dello Smart Social Sensor esporrà delle interfacce RESTful consumate dal layer di Data Visualization.

Il layer di Data Visualization si occupa invece dell'estrazione dei risultati delle analisi e della scelta della metafora di visualizzazione migliore per fornire in tempo reale i risultati delle elaborazioni. Questo layer si baserà su tecnologie open source che sfruttano il supporto dato dai moderni browser nell'erogazione di grafica avanzata, usando lo standard SVG.

La Figura 6 mostra l'architettura di uno Smart Social Sensor:

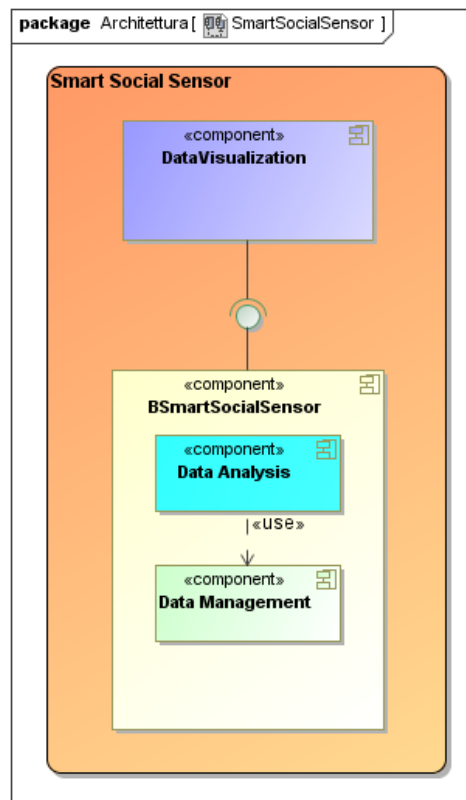


Figura 6. Architettura di uno Smart Social Sensor

Dalla Figura 6 risulta evidente come uno Smart Social Sensor sia trasversale ai componenti dell'architettura della piattaforma di monitoraggio descritti in Figura 4. Uno Smart Social Sensor sarà sempre caratterizzato da:

- un obiettivo di generazione di conoscenza;
- una sorgente di Social Big Data;
- un set di processi di Data Management;
- un set di processi di Data Analysis;
- un set di processi di Data Visualization;
- un template di parametri di configurazione per le tre tipologie di processi descritti.

I processi di Data Management e Data Analysis saranno implementati da un micro-servizio, mentre quelli di Data Visualization dal componente di GUI.

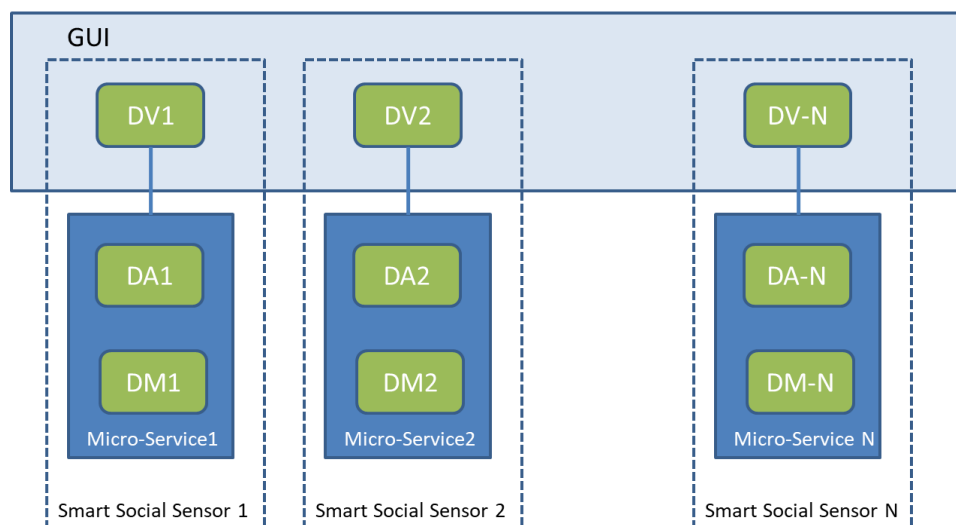


Figura 7. Relazione fra Smart Social Sensor e micro-servizi

6. Data requirements per la realizzazione del prototipo di sistema di monitoraggio

Per la realizzazione del prototipo di sistema di monitoraggio dei canali social all'interno del progetto EasyPAL, è necessario individuare e caratterizzare le sorgenti di dati. E' richiesto quindi per ogni sorgente dati:

- la caratterizzazione della sorgente in termini di tipologia di Social Network (es. Facebook, Twitter);
- l'identificazione della pagina, gruppo o utente da monitorare;
- l'identificazione di termini o parole chiave da monitorare;
- le credenziali di autenticazione per l'accesso alle API della sorgente.

Un'attività di questo tipo è propedeutica all'identificazione dell'obiettivo di analisi e permetterà di dare valore alle analisi stesse che il prototipo produrrà, potendo generare informazioni e conoscenza a partire da dati reali e con caratteristiche congrue al paradigma dei Big Data.

Bibliografia

- Agarwal, R., Dhar, V. (2014). "Editorial – big data, data science, and analytics: the opportunity and challenge for is research". *Information Systems Research*, 25 (3), pp. 443–448
- Bello-Orgaz G, Jung JJ, Camacho D. (2016) "Social big data: recent achievements and new challenges". *Inf Fusion*.
- Bollen J., Mao H., Zeng X., (2011) "Twitter mood predicts the stock market," *Journal of Computational Science*, vol. 2, no. 1.
- Burgess J, Bruns A. "Twitter archives and the challenges of "big social data" for media and communication research". *M/C J*. 2012.
- Coté M. "Data motility: the materiality of big social data". *Cult Stud Rev*. 2014.
- Culotta A., (2010) "Towards detecting influenza epidemics by analyzing twitter messages," in *Proceedings of the First Workshop on Social Media Analytics (SoMA)*.
- De Mauro, A., Greco, M., & Grimaldi, M. (2016). "A formal definition of Big Data based on its essential features". *Library Review*, Vol. 65, No.3, pp.122-135.
- Elia, G., Polimeno, G., Solazzo, G., Passiante, G. (2018) "A Creating Value for organizations through Big Data: a Framework based on a systematic literature review", *Proceedings of IFKAD 2018 - International Forum on Knowledge Asset Dynamics on "Societal Impact of Knowledge and Design"*, 4-6 July 2018, Delft, Netherlands.
- Gandomi A, Haider M. Beyond the hype: big data concepts, methods, and analytics. *Int J Inf Manag*. 2015;35(2):137–44.
- Guellil I, Boukhalfa K. Social big data mining: a survey focused on opinion mining and sentiments analysis. In: *2015 12th international symposium on programming and systems (ISPS)*. New York: IEEE; 2015. p. 1–10.
- Guo, B., Yu, Z., Zhou, X., Zhang, D. (2014). "From participatory sensing to mobile crowd sensing". In *Pervasive Computing and Communications Workshops (PERCOM Workshops)*, IEEE.
- Housley W, Procter R, Edwards A, Burnap P, Williams M, Sloan L, Rana O, Morgan J, Voss A, Greenhill A. (2014) "Big and broad social data and the sociological imagination: a collaborative response". *Big Data Soc.*
- Ishikawa H. "Social big data mining". Boca Raton: Taylor & Francis Group, CRC Press; 2015.
- McAfee, A. & Brynjolfsson, E. (2012). "Big data: the management revolution". *Harvard Business Review*, 90(10), 61-67
- Nguyen DT, Hwang D, Jung JJ. "Time–frequency social data analytics for understanding social big data". Cham: Springer; 2015.
- Olshannikova, E., Olsson, T., Huhtamäki, J., & Kärkkäinen, H. (2017). Conceptualizing big social data. *Journal of Big Data*.
- Pentland A. "Social physics: how good ideas spread-the lessons from a new science". New York: The Penguin Press, Penguin Group; 2014.

- Quercia D., Ellis J., Capra L., Crowcroft J., (2011) "In the Mood for Being Influential on Twitter," in Proceedings of the 3rd IEEE International Conference on Social Computing (SocialCom).
- Sakaki T., Okazaki M., Matsuo Y., (2010) "Earthquake shakes Twitter users: Real-time event detection by social sensors". 19th International Conference on World Wide Web, Raleigh, NC, USA, April 26-30.
- Sivarajah, U., Kamal, M. M., Irani, Z., & Weerakkody, V. (2017). Critical analysis of Big Data challenges and analytical methods. Journal of Business Research, 70, 263-286.
- Tang J, Chang Y, Liu H. "Mining social media with social theories: a survey". ACM SIGKDD Explor Newsl. 2014.
- Vasudevan V., Wickramasuriya J., Zhao S., Zhong L., (2013) "Is Twitter a good enough social sensor for sports TV?" 2013 IEEE International Conference on Pervasive Computing and Communications Workshops, San Diego, CA, USA, March 18-22,.
- Wamba, S. F., Akter, S., Edwards, A., Chopin, G., & Gnanzou, D. (2015). "How 'big data' can make big impact: Findings from a systematic review and a longitudinal case study", International Journal of Production Economics, Vol. 165, pp. 234-246.
- Yoon, S., Lee, J., Kim, H. (2016). Research into a Visualization Analysis of Bigdata for the Decision Making of a Tourism Policy.