UNIVERSITÀ DI PISA

Corso di Laurea in Informatica Umanistica

RELAZIONE

Studio della diffusione delle news sul social network Facebook

RELATORI:

Prof.re Andrea Marchetti

Prof.re Mirko Tavosanis

Ing.re Davide Gazzè

CANDIDATO:

Andrea Tonacchera

Anno accademico 2014/2015

Indice

1. Introduzione

Oggigiorno i social network fanno parte della nostra vita; essi permettono di condividere opinioni, foto, video e news oltre che a creare nuove relazioni tra persone.

Infatti, attraverso la creazione di un profilo personale chi si iscrive può raccontare qualcosa di sé; pubblicare immagini, link, musica e video; partecipare a gruppi tematici e alle relative discussioni; interagire con altri utenti in vari modi.

I social network si possono considerare come delle piazze virtuali che espandono la nostra possibilità di comunicare trasformandoci in agenti attivi di campagne a favore di quello in cui crediamo. Il social network ha totalmente aperto e modificato le frontiere della comunicazione digitale integrando in un solo "contenitore" vari servizi: il profilo dell'utente, il blog, la messaggistica, il download della musica, la galleria fotografica, la community.

Nei social network, inoltre, è esaltata una delle caratteristiche chiave del web, cioè la partecipazione, l'interesse attivo dei membri a trovare amici e coltivare relazioni.

La Community è costituita da un gruppo di utenti che si aggrega in base a interessi comuni, per scambiarsi informazioni, cercando il confronto attraverso varie modalità di interazione.

Al giorno d'oggi, i Social Network sono utilizzati anche per la condivisione di notizie provenienti della varie testate giornalistiche online. Infatti, all'interno di un Social Network, l'utente può esprimere una preferenza, commentare o semplicemente condividere una notizia. Ognuna di queste notizie appartiene a una categoria (cronaca, politica, sport, etc.) definita dalla testata giornalistica stessa. L'analisi di queste informazioni forniscono indicazioni sul grado di interesse di una notizia e della relativa categoria di appartenenza (cronaca, sport, politica, attualità), fornendo importanti informazioni circa le preferenze della popolazione.

Con la diffusione dei personal computer e l'implementazione delle piattaforme informatiche le notizie vengono diffuse in tempo reale online.

Per questa motivazione, la quantità dell'informazione disponibile sta crescendo vertiginosamente, con i media sociali, internet, la digitalizzazione e la crescita della velocità di accesso ai contenuti che si trovano in rete.

Lo scopo della tesi è di studiare la diffusione delle notizie pubblicate dalle più importanti testate giornalistiche italiane (La Stampa, il Corriere della Sera, La Repubblica, il Sole 24 Ore, etc.) sulla piattaforma Facebook. Il sistema prevede lo sviluppo di diverse parti: acquisizione, analisi e visualizzazione. Per l'acquisizione saranno presentati due moduli di raccolta dati. Il primo modulo raccoglie i link delle notizie provenienti dalle varie testate (tramite l'uso dei loro Feed RSS), mentre il secondo cattura la diffusione di ogni notizia su Facebook in termine di numero di like, commenti, share e click. Lo strumento di analisi creato ha il compito di estrarre dalle informazioni raccolte le statistiche che ne derivano come: il grado di interesse della popolazione per la testata giornalistica e per la categoria consultata. Inoltre sono stati fatti degli studi sul confronto di testate appartenenti a gruppi ad interesse specifico (sport, economia, ambiente) oppure generici e sul grado di diffusione di una singola testata.
In generale questa tesi ha l'obbiettivo di rispondere alle seguenti domande:

- Quante notizie non ricevono diffusione su Facebook?
- Quali sono le testate giornalistiche online più attive?
- Quali sono le testate giornalistiche online che ricevono più attenzione su Facebook? Dipende dalla taratura o dalla popolarità della testata?
- Qual è l'andamento delle categorie nel tempo?
- Che tipo di notizie si diffondono in Italia?
- Ricevono più interazioni su Facebook le testate generalistiche (La Stampa, il Fatto Quotidiano) o quelle specialistiche (la Gazzetta dello Sport, il Sole 24 Ore)?

I giornali online pubblicano una grande quantità di notizie**,** utili per lo studio dell'opinione pubblica**.** Limitatamente alle nostre conoscenze, non esiste alcun strumento che provveda alla memorizzazione delle statistiche di notizie e della relativa diffusione sui social network (Facebook in particolare).
Ci sono strumenti shareware [1] attualmente disponibili che analizzano queste statistiche a livello di post pubblicato sul social network (es. facebook) estrapolando

[1] Lo shareware, conosciuto anche come trial, è la prova di un software che dopo un limite di tempo sarà a pagamento

le statistiche d'interesse per contenuti specifici (una parola e tutti i collegamenti ad essa riferiti).

L'obiettivo del lavoro è quello di creare uno strumento che basandosi sui link dei Feed RSS, pubblicati sulle singole testate giornalistiche online, analizzi i contenuti di tutte le notizie e ne estrapoli le relative informazioni.

Lo strumento si può estendere ad altri social network anche se questo lavoro si è concentrato esclusivamente su Facebook perché con le API l'acquisizione dei dati è più semplice oltre al fatto che Facebook essendo il social network più diffuso in Italia è anche fonte di campioni eterogenei.

2. Stato dell'arte

Esistono software con cui è possibile rimanere aggiornati in maniera rapida attraverso dei lettori di notizie online (chiamati feed reader) che analizzano ed estrapolano i contenuti presenti nel feed (es: Readly, FeedDemon e QuiteRSS).
Inoltre, sono disponibili in rete diversi strumenti che permettono il social media monitoring attraverso l'analisi dei post pubblicati all'interno di un social network.
Alcuni di questi strumenti utilizzati vengono sotto descritti: Trackur, SocialMention, CyberAlert, Hootsuite, SocialMediaItalia, Moreover Neewdesk, Social trends, KPI6**,** Flipboard**.**

2.1 Trackur

Trackur [2] è un social media monitoring tool progettato per monitorare il livello di gradimento di una azienda e del marchio dei prodotti ad esso collegati. Scansiona centinaia di milioni di pagine, notizie inclusi blog, news, forum, video, immagini.
Questo tool controlla le principali tendenze dei media e dei social media: blog, social networks, twitter, facebook. Gli articoli e oggetti vengono trovati quasi in tempo reale. Tramite i grafici generati è possibile vedere l'andamento delle opinioni.

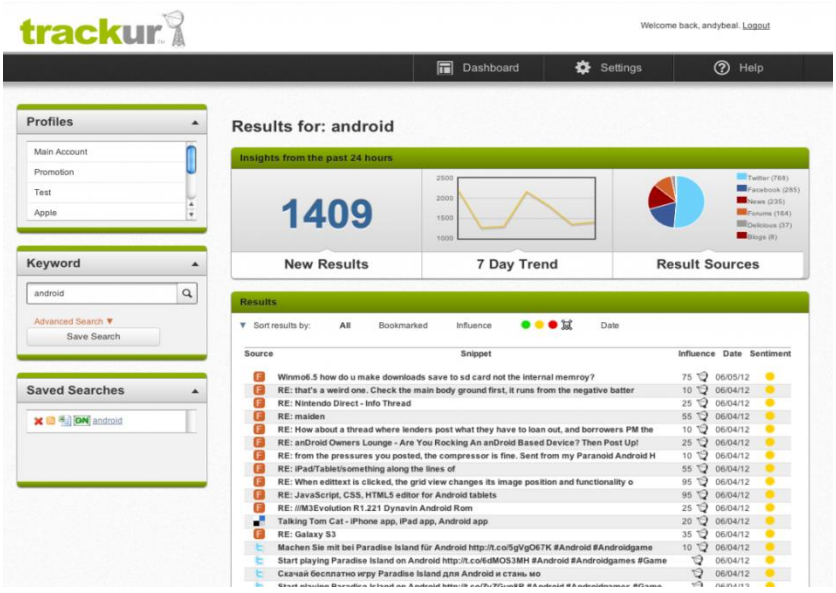


Figura 2.1: Interfaccia Trackur

[2] http://www.trackur.com

2.2 Socialmention

Social Mention[3] è un tool di social media monitoring dall'interfaccia semplice; questo strumento è utilizzato per la ricerca e l'analisi, in tempo reale, di parole chiave all'interno di vari siti web, ad esempio: Facebook, Yahoo Answer, Youtube, Photobucket ecc…

Social Mention, a seconda della parola chiave digitata, restituisce i risultati, il numero di opinioni positive, neutrali e negative, l'ultima volta in cui la parola ricercata è stata usata.

Questo tool permette anche di identificare le parole chiave più utilizzate, gli users più attivi, gli hashtags(nota) più utilizzati e nei siti analizzati quante volte questa parola compare.

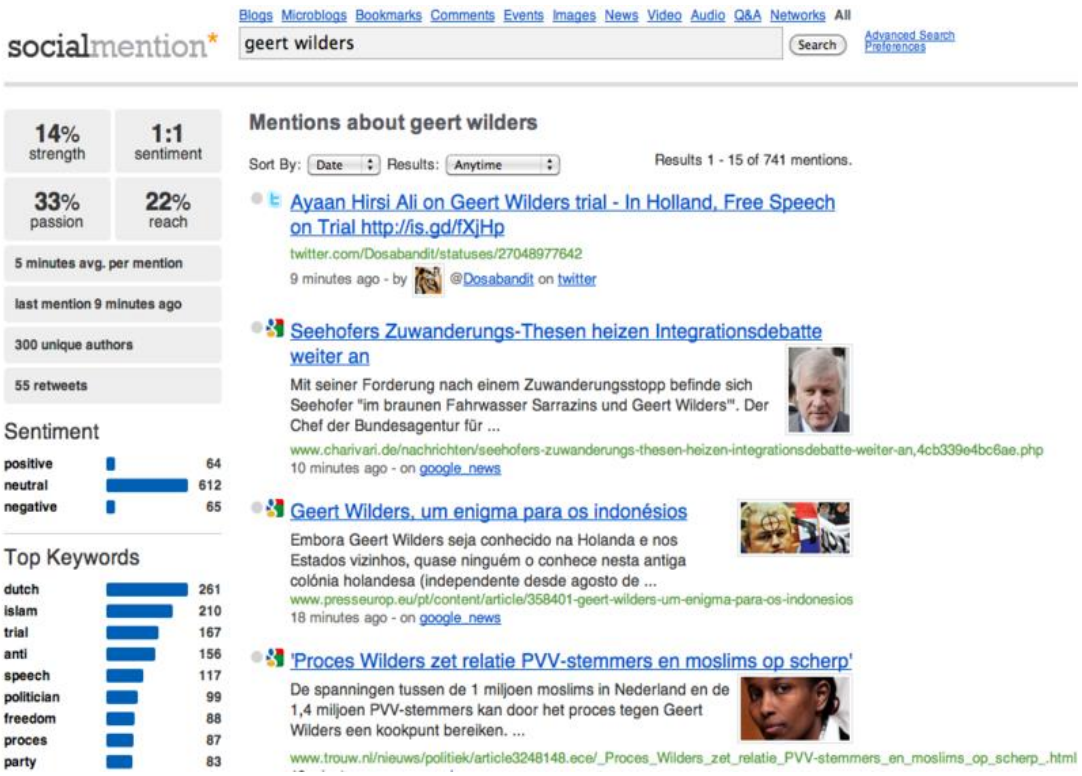
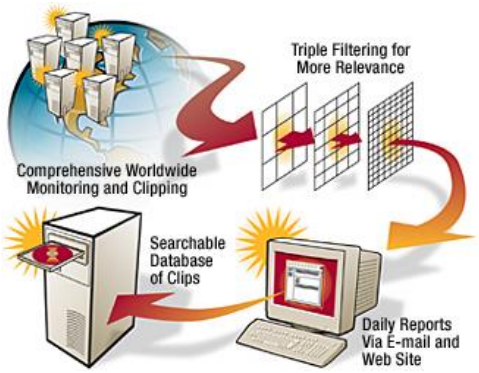


Figura 2.2: Interfaccia Socialmention

[3] http://socialmention.com/

2.3 CyberAlert

CyberAlert[4] offre un servizio "tutto in uno" di monitoraggio dei social media dei quali ha un ampia copertura, servizio di allarme email giornaliero e archivio online.



I dati estrapolati con questo strumento sono misurati accuratamente e da la possibilità di trovare e catturare quello che si ritiene più necessario alle proprie esigenze.

Mentre alcune organizzazioni utilizzano personale per monitorare le notizie e i social media, la maggior parte delle aziende e agenzie utilizzano sistemi come Cyberalert.

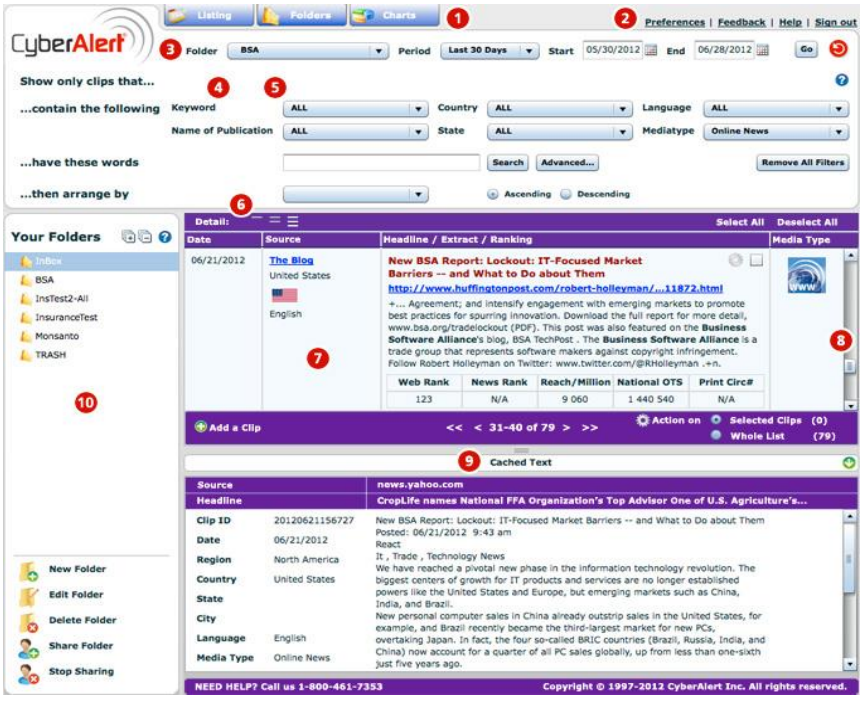


Figura 2.3: Interfaccia Cyberalert

[4] http://www.cyberalert.com

2.4 Hootsuite

Hootsuite Syndicator Pro[5] è un tool che permette di visualizzare in modo semplice e veloce Feed RSS su un'unica dashboard e condividerli sui social media. È possibile importare i Feed RSS del proprio blog, ed impostare la possibilità di pubblicare automaticamente gli aggiornamenti ai nuovi post direttamente su Twitter o Facebook.

Si utilizza per esportare i feed esistenti da altri servizi via OPML o file XML, o aggiungere nuovi Feed RSS individualmente.

Con Hootsuite è anche possibile importare i Feed RSS del proprio blog, ed impostare la possibilità di pubblicare automaticamente gli aggiornamenti ai nuovi post direttamente su Twitter o Facebook.

Figura 2.4: Interfaccia Hootsuite

[5] https://hootsuite.com/

2.5 Morover Newdesk

È uno strumento efficace nel trovare notizie su social media, stampa, TV e la copertura radio in un unico servizio di monitoraggio.

Newsdesk[6] aggiunge 2.500 nuovi post al minuto e consente di analizzare oltre 250 milioni di articoli in breve tempo.

Crea relazioni dettagliate per la valutazione dei media con uno dei database multimediali più completi al momento.

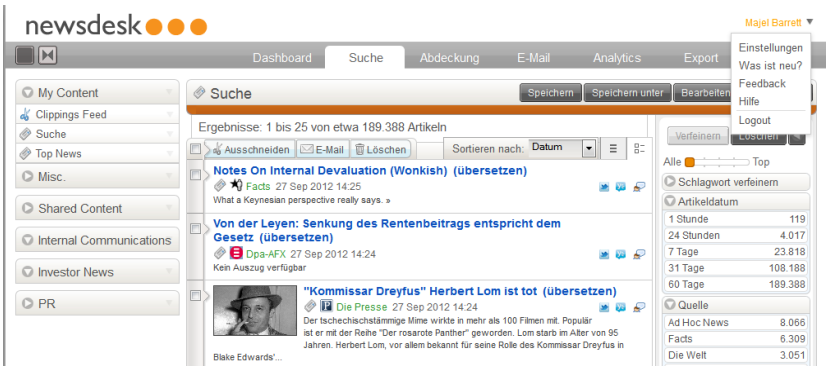


Figura 2.5: Interfaccia Newdesk

Il Team di marketing è in grado di monitorare contemporaneamente più argomenti e parole in tutte le lingue e mercati internazionali, per una completa visione a 360 gradi del panorama dei media.

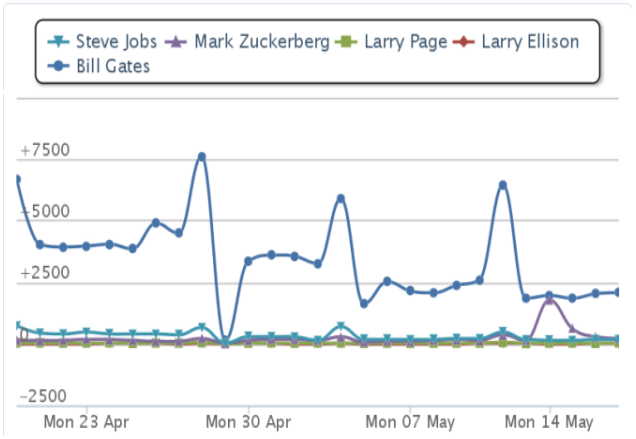
Offre una soluzione di monitoraggio dei media conveniente, con clienti come professionisti di marketing e agenzie di Public Relation con liste di clienti globali. Cattura eventi che avvengono in tempo reale, con copertura mediatica in tutto il mondo condividendo anche i risultai con dei grafici e statistiche giornaliere/mensili.

Figura 2.5.1: Interfaccia Newdesk

[6] http://www.moreover.com/newsdesk/media-monitoring

2.6 Social Trends

SocialTrends [7] ha come obiettivo principale l'analisi del'attività, popolarità e influenza di personaggi famosi, quotidiani, partiti politici, ecc. sui Social Media più popolari.



I soggetti analizzati sono suddivisi in categorie (quotidiani, politici, gionalisti, ecc.) e per ognuna di esse vengono mostrati i "***top posts***", i "***top tweets***" e i "***top videos***". Queste metriche sono calcolate usento rispettivamente i posts, tweets e video di maggiore successo negli ultimi 15 giorni.

SocialTrends è un progetto del gruppo **Web Application for Future Internet** dell'Istituto di Informatica e Telematica del CNR di Pisa ed opera su dati di pubblico dominio messi a disposizione da Facebook, Twitter e YouTube.

Figura 2.6: Interfaccia Socialtrends

[7] http://www.social-trends.it/home

2.7 KPI6

KPI[8] 6 è una piattaforma, ancora in via di sviluppo, di Network Analysis che permette l'ascolto e il monitoraggio delle conversazioni online e la raccolta di dati ed insight sui propri competitors.

Nella fase iniziale di lancio, KPI 6 si concentra sulle attività di listening della rete e sulla social network analysis come: Facebook, Twitter, Google Plus, Youtube Google Analytics.

Il principale obiettivo è quello di monitorare le conversazioni spontanee che avvengono sui social network offrendo la possibilità di estrapolare dati (quantitativi/qualitativi) ed insight utili a implementare strategie di web e social media marketing sempre più efficaci.

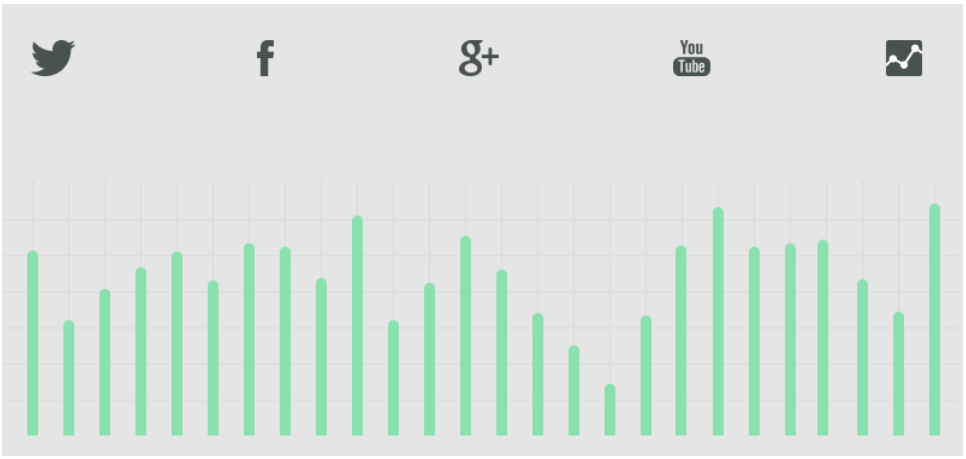


Figura 2.7: Immagine relativa al listening KPI6

[8] https://kpi6.com/

2.8 Flipboard

Flipboard[9] è un'applicazione definita, dagli stessi creatori, una rivista sociale, ossia un magazine personalizzato progettato principalmente per telefoni e tablet, alla quale è tuttavia possibile accedere anche da PC.

Funziona inoltre su diversi sistemi operativi, come Android, iOS e windows phone. In iOS si apre come una normale applicazione, ma su Android è possibile configurarla come un widget se lo si desidera.

Flipboard, ha una grafica intuitiva e attraente.

In base al tipo di notizie sulle quali vogliamo essere aggiornati, si può cliccare sul quadrato corrispondente che ci permette di seguire le notizie relative alla categoria selezionata.

Si apre una schermata dove compaiono le notizie della categoria scelta e cliccandoci troveremo moltissime immagini, video e articoli da sfogliare.

Cliccando su Facebook o un altro social network è possibile visionare ciò che i nostri amici hanno pubblicato, i video e tutti i post dei vari gruppi ai quali siamo iscritti.

Flipboard raccoglie infatti i contenuti dei social network, news, pubblicazioni e blog in un formato che ricorda una rivista.

Gli utenti hanno la possibilità di creare le proprie rassegne, nonché sottoscrivere riviste di altri utenti o brand.

Effettuare la propria rivista non è essenziale, ma può essere divertente e facile. Ogni articolo di contenuti che visualizziamo in Flipboard può essere aggiunto a una rivista oppure modificato.

[9] https://flipboard.com/

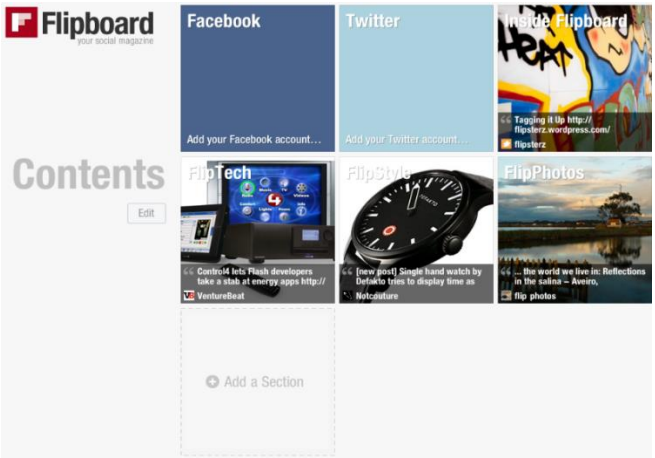


Figura 2.8: Immagine Interfaccia Flipboard

3. Design Architetturale

Il sistema è composta da diversi moduli: acquisizione, analisi e visualizzazione.

- I moduli di acquisizione effettuano la raccolta dati. Il primo modulo raccoglie i link delle notizie provenienti dalle varie testate (tramite l'uso dei loro Feed RSS), mentre il secondo cattura la diffusione di ogni notizia su Facebook in termine di numero di like, commenti, share e click.
- I moduli di analisi hanno il compito di estrarre le informazioni dai dati raccolti mentre le analisi sono state condotte su vari livelli: news, testata giornalistica e categoria.
- La visualizzazione presenta i dati raccolti e le analisi effettuate.

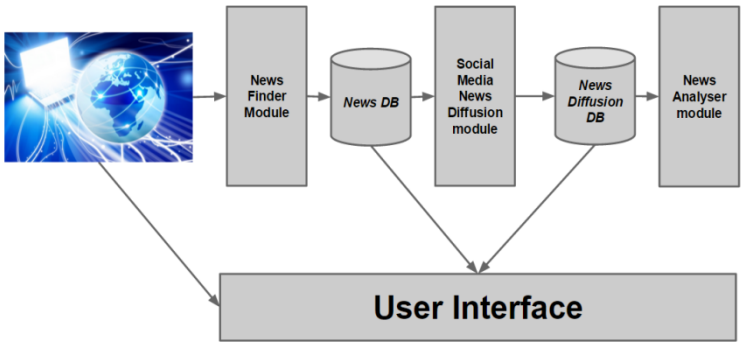


Figura 3: Design architetturale

3.1 Feed Rss

Con i termine Feed Rss intendiamo un flusso di informazioni in formato XML che permette la distribuzione di contenuti sul Web.

Il motivo principale per cui sono famosi i Feed Rss (acronimo di Really Simple Syndication) è la semplicità con cui permettono agli utenti di rimanere sempre aggiornati sugli articoli di un sito in continuo aggiornamento.

Infatti, la natura del Feed Rss è strettamente legata a quella di un blog, uno spazio web che muta continuamente e che deve essere seguito con attenzione.

Grazie ai Feed Rss è possibile rimanere sempre aggiornato, in maniera del tutto gratuita, sulle ultime pubblicazioni blog e siti web con grande affluenza di informazioni: basta un semplice feed reader che permette di leggere il nuovo articolo senza dover raggiungere l'indirizzo del blog in questione.

Un'applicazione in grado di interpretare un documento ne effettua il parsing (lettura), ovvero una scansione del documento che individua i tag e isola i diversi elementi, per poi convertire i contenuti decodificati nel formato utile all'obiettivo: ad esempio un feed reader può estrarre i titoli di tutti gli elementi (chiamati item) per visualizzare la lista degli articoli di un giornale online, mentre un aggregatore Web può estrarre i contenuti del flusso per convertirli in linguaggio HTML e incorporarli all'interno delle proprie pagine.

I Feed Rss sono stati presi dalle seguenti testate giornalistiche: Repubblica, Corriere della sera, Il fatto Quotidiano, Il Tirreno, Ansa, La Nazione, La Stampa, Il Messaggero, Il Giornale, Libero, Gazzetta Dello Sport, Il Sole 24 Ore.

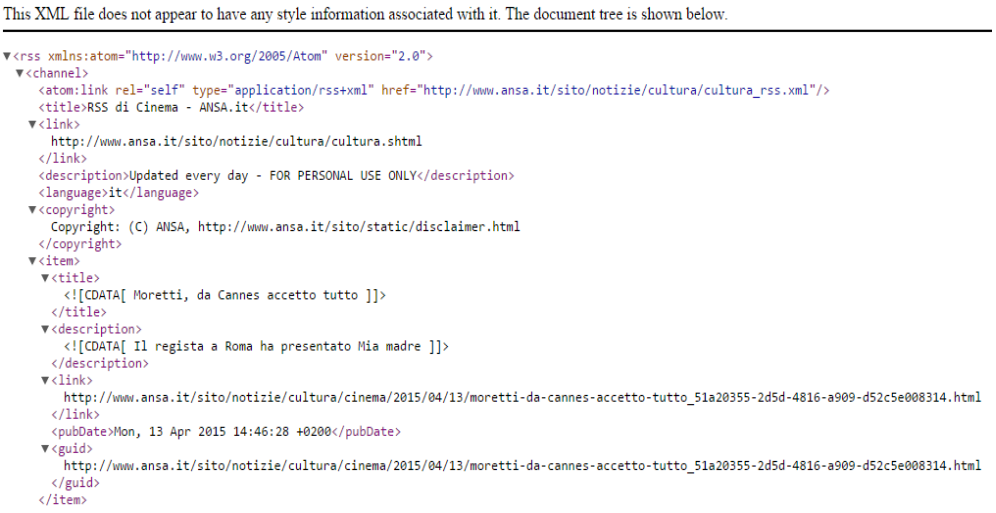


Figura 3.1: Un esempio di file xml

3.2 Struttura Database

Il database è composto da varie tabelle:

- **Blacklist:** contenente tutti gli url da scartare in fase di analisi perché influenzano la qualità delle stesse. Esempi di url in blacklist sono quelle delle singole testate giornalistiche (es: www.ilcorriere.it, www.gazzetta.it).
- **News_categories:** contenente tutti gli url delle notizie estratti dal Feed Rss con le rispettive categorie.
- **Categories:** contiene tutte le categorie estratte dai feed rss che vengono inserite in un unica categoria, visto che le testate giornalistiche hanno le stesse categorie ma chiamate diversamente (es: Energia, Animali in Ambiente).
- **Link_diffusion: dove viengono salvate i valori di dissusione su Facebook** t(es: click_count, like_count, share_count, normalized_url, url, comment_count, date) a livello di singola notizia.
- **Link_history:** questa tabella contiene le stesse informazioni di link_diffusion e l'infomazione temporale dell'istante di raccolta di suddette informazioni.
- **Log:** si trova il resoconto delle giornate di crawling, di ora in ora il numero di notizie sono state inserite o aggiornate.
- **News:** tabella contenente le notizie che vengono catturate dal Feed Rss con: titolo, link, autore, descrizione, data di pubblicazione, data di crawling.
- **Newspapers:** tabella contenente le testate giornalistiche online divise per categoria con il relativo link Rss da dove crawlare le notizie.
- **Parameters:** tabella contenete i parametri della piattaforma. Attualmente sono presenti due parametri, maxAgeFeed e maxAgeFeedFB. Rispettivamente indicano i tempo di aggiornamento delle notizie e della diffusione da Facebook.

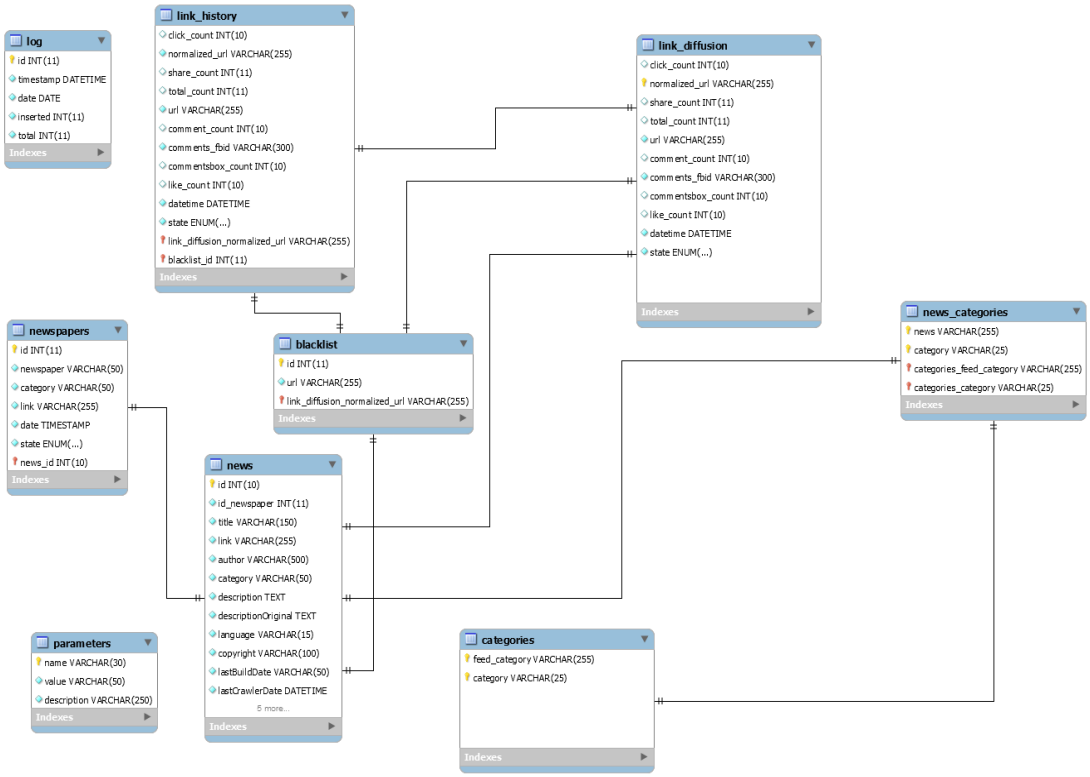


Figura 3.2: Come è strutturato il database

3.3 Bootstrap

Bootstrap è una serie di elementi e funzioni per il web-design personalizzabili raccolti in un unico strumento. Quando progettano un sito web con Bootstrap, gli sviluppatori possono scegliere quali elementi utilizzare. Cosa ancora più importante, gli sviluppatori possono avere la certezza che gli elementi scelti non saranno in conflitto tra di loro. È come avere un puzzle in cui tutti i pezzi si incastrano perfettamente, a prescindere da quello scelto.

Gli elementi personalizzabili contenuti in Bootstrap sono una combinazione di HTML, CSS e JavaScript. Grazie alla comodità dell'open-source, Bootstrap viene migliorato continuamente.

Questa libreria è facile da usare poiché qualsiasi utente può decidere se creare un nuovo sito html con dei demo già pronti che offre, o ripartire da zero utilizzando le classi che un utente preferisce. Inoltre è possibile in pochissimi passi rendere il

proprio sito web responsive[10] poiché include i CSS media query, disponendo di un ottima guida d'utilizzo nella sua homepage se l'utente ne avesse bisogno.

3.4 Jquery

Per lo svilippo è stato usato jQuery, un framework sviluppato da John Resig a partire dal 2006 con il preciso intento di rendere il codice più sintetico e di limitare al minimo l'estensione degli oggetti globali per ottenere la massima compatibilità con altre librerie.

Da questo principio è nata questa libreria in grado di offrire un'ampia gamma di funzionalità, che vanno dalla manipolazione degli stili CSS e degli elementi HTML, agli effetti grafici per passare a comodi metodi per chiamate AJAX cross-browser. Il tutto, appunto, senza toccare nessuno degli oggetti nativi JavaScript.

Nel progetto viene usato per richiamare le API tramite chiamata AJAX per la creazione di grafici interattivi (si veda il codice sotto).

```
function esempio() {
    $.ajax({
        dataType: "json",
        url: "./api/getActiveOnlTestate.php",
        type: "POST",
        data: {'field': type},
        success: function(data) {
            grafo3(data);
            $("#loading").hide();
        },
        error: function(e) {
            console.log(e.responseText);
        }
    });
}
```

[10] Si adatta ad ogni tipo di dispositivo: mobile o non.

3.5 Highcharts/Highstock

Highcharts (Javascript Charting Library) è una libreria per la realizzazione di grafici scritta in linguaggio JavaScript, che offre un modo semplice di aggiungere grafici interattivi ai propri siti o applicazioni web.

Attualmente supporta i tipi di grafico a linea, spline, area, areaspline, column, bar, pie, scatter, angular gauges, arearange, areasplinerange , columnrange e polar.

Funziona su tutti i moderni browser, inclusi iPhone/iPad ed Internet Explorer dalla versione 6. Il web attualmente gira tutto intorno agli standard aperti e a tecnologie non proprietarie.

La libreria Highcharts, che usa il linguaggio SVG, rientra perfettamente in tale contesto.

Il punto di forza di Highcharts è la selezione automatica della migliore modalità di rendering. Inoltre, come già detto, questa libreria rende superato l'uso di Flash per la grafica vettoriale.

Una delle caratteristiche chiave di Highcharts è la disponibilità sotto diverse licenze, free o meno. L'uso della libreria è libero per siti personali o di organizzazioni no-profit. Questo permette modifiche personali ed una grande flessibilità.

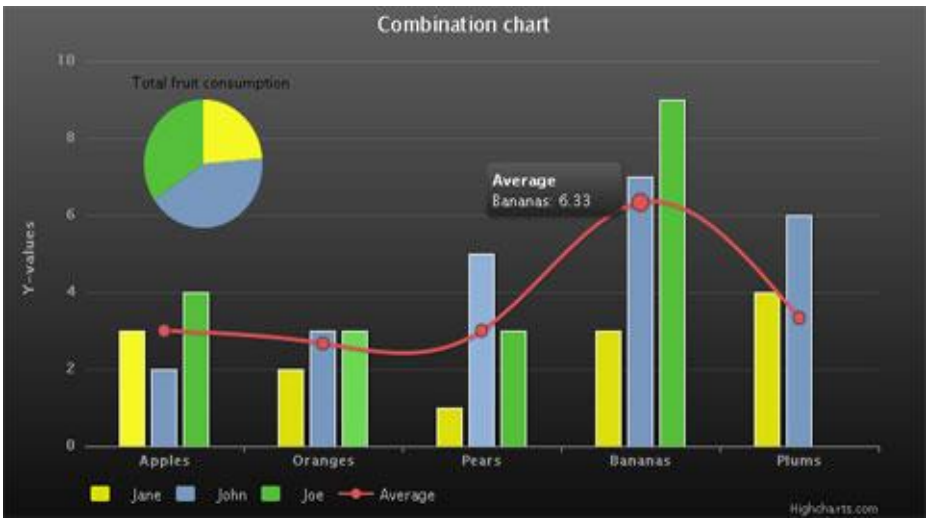


Figura 2.7: Esempio di Highcharts

3.6 XML Parser

In generale un parser è un software in grado di effettuare l'analisi sintattica di un testo scritto in un determinato linguaggio. Nello specifico, un parser XML è un software in grado di effettuare l'analisi sintattica di un documento XML.
La libreria utilizzata per lo sviluppo di questa piattaforma è SimpleXml. Tale libreria ha il grosso vantaggio di avere un'interfaccia ad oggetti molto semplice ed intuitiva, che, a differenza del più prolisso DOM, richiede pochissime righe di codice per accedere agli elementi interessati e mantiene intatta la struttura del file XML anche nella sua rappresentazione interna (a differenza di SAX che ci obbliga a tener traccia manualmente delle gerarchie durante la gestione degli eventi generati).
Questa libreria è stata utilizzata, creando uno script in PHP che esegua il parsing del Feed RSS 2.0 per l'estrazione di tutte le informazioni annesse all'url del Feed Rss.
Questo ha portato alla creazione di un file nominato newsReader.php con il qual catturiamo le notizie.

Esempio 3.5: Un esempio di codice che mostra come usare spimplexml.

```
$feed_url ='Esempio/cronaca_rss.xml';
$feed = simplexml_load_file($feed_url);
foreach($feed->channel->children()  as $type => $item) {
    if($type == "item"){
        $title = "";
        if(isset($item->title)) {
            $title = $item->title;
            $title = strip_tags($title);
        };
    };};
```

3.7 Facebook FQL Query

Facebook Query Language (FQL) è un linguaggio, molto simile a SQL, creato da Facebook per interrogare le vaste base dati in maniera facile.
La chiamata si divide in 2 fasi:

1. L'esecuzione della query;
2. L'attesa dei risultati

FQL è molto utile perchè permette di utilizzare funzionalità avanzate che con le Graph API (altro sistema per estrarre informazioni) sarebbero impossibili, ad esempio inserire chiamate multiple in una singola chiamata.
È stato utilizzato per passare l'url della notizia, scaricato precedentemente dal Feed Rss, a Facebook.

Esempio 3.6: Un piccolo estratto di codice che invia l'url della notizia selezionata a Facebook.

```
function fqlQuery($links, $api, $secret) {
     $urls = "'".implode("','", $links)."'";
     $fql = "SELECT url, click_count, comment_count, comments_fbid,
commentsbox_count, like_count, normalized_url, share_count, total_count FROM
link_stat WHERE url IN ($urls)";
$apifql=https://api.facebook.com/method/fql.query?format=json&query=.urlenco
de($fql)."&access_token=$api|$secret";
     $fb_json = file_get_contents($apifql);
     return json_decode($fb_json);
}
```

La funzione file_get_contents legge il contenuto dell'URL definito e lo restituisce sotto forma di stringa.
L'utilizzo più frequente delle funzione prevede un unico parametro (l'unico obbligatorio) il quale consiste nell'indicazione del file da leggere:
$miastringa = file_get_contents('/cartella(o url)/file.txt').

Questa operazione è possibile a condizione che le impostazioni del web-server non la vietino. Ovviamente l'accesso a dei file remoti comporta dei rischi e pertanto è consigliabile farvi ricorso solo qualora si sia veramente certi del contenuto del file remoto che si desidera leggere.

3.8 Moduli di raccolta dati

Per fare queste raccolte dati utilizziamo determinati software di raccolta dati (denominato crawler).
Un crawler è un particolare tipo di bot (abbreviazione di robot, in Internet è utilizzato per indicare un programma o uno script in grado di automatizzare processi e comandi tipicamente eseguiti da esseri umani) utilizzato dai motori di ricerca per analizzare ed indicizzare i contenuti pubblicati in rete, e da qua deriva il termine italianizzato "crawlato, crawlare".
È un software automatizzato che principalmente scarica il contenuto delle pagine web e si autoalimenta individuando, da ogni pagina, un elenco di ulteriori URL da analizzare.
I moduli di raccolta servono a prendere dati e si dividono in 2 fasi:

1. **Raccolta dati attraverso Feed rss:** il feed rss attraverso uno script, che spiegheremo di seguito, viene analizzato ed crawlato.
2. **Raccolta dati attraverso Facebook:** a Facebook vengono mandate delle richieste attraverso uno script contenete il linguaggio FQL Query.

3.9 Feed Rss Crawler

È stato creato uno script(schedulerNews.php) in PHP che contiene tutti i comandi. Partendo dalla verifica dei parametri di connessione analizza i tempi di aggiornamento dei singoli feed, analizza il contenuto del feed (url) verificando se esistono campi specifici come definiti nello script:

- title;
- link(url della notizia);
- author;
- category;
- datatime(quando è stato pubblicato il feed);
- description(descrizione della notizia);
- language;
- copyright;
- lastBuildDate(data pubblicazione notizia);

Esempio 3.8: Codice che mostra quali tag andare ad estrarre dal Feed Rss che gli viene passato.

```
$feed = simplexml_load_file($feed_url);
   foreach($feed->channel->children()  as $type => $item) {
       if($type == "item"){
           $title = "";
           if (isset($item->title)) {
               $title = $item->title;
               $title = strip_tags($title);
           }
           $link = "";
           if (isset($item->link)) {
               $link = $item->link;
               $link = strip_tags($link);
           }
           $author = "";
           if (isset($item->author)) {
               $author = $item->author;
```

```
                $author = strip_tags($author);
            }
            $category = "";
            if (isset($item->category)) {
                $category = $item->category;
                $category = strip_tags($category);
            }
            $dataTime = "";
            if (isset($item->pubDate)) {
                $dataTime = $item->pubDate;
                $dataTime = strip_tags($dataTime);
            }
            $description = "";
            if (isset($item->description)) {
                $description = $item->description;
                $description = strip_tags($description);
            }
            $language = "";
            if (isset($item->language)) {
                $language = $item->language;
                $language = strip_tags($language);
            }
            $copyright = "";
            if (isset($item->copyright)) {
                $copyright = $item->copyright;
                $copyright = strip_tags($copyright);
            }
            $lastBuildDate = "";
            if (isset($item->lastBuildDate)) {
                $lastBuildDate = $item->lastBuilDate;
                $lastBuildDate = strip_tags($lastBuildDate);
            }
        }
    }
```

Completata l’analisi i vengono estrapolati ed inseriti nel database.

Esempio 3.8.1: Codice che mostra cosa e dove andare a inserire le informazioni.

```
$sql_news = "INSERT INTO news (id_newspaper, title, link, author, category,
description, language, copyright,  lastBuildDate, lastCrawlerDate, date)
VALUES('$id_newspaper','$title','$link','$author','$category','$description'
,'$language' ,'$copyright','$dataTime','$dataTimeInsert','$date') ON
DUPLICATE KEY UPDATE category = '$category',
lastCrawlerDate='$dataTimeInsert'";
$query = mysqli_query($conn, $sql_news)or die(mysqli_error($conn));
```

Una volta completato l'inserimento dei dati si conclude con il comando exec() che esegue l'istruzione passata da $command: la funzione non invia nessun output, restituisce semplicemente l'ultima linea dal risultato del comando.
Terminato, lo stato nel database si aggiorna cambiandolo da downloading a complete.

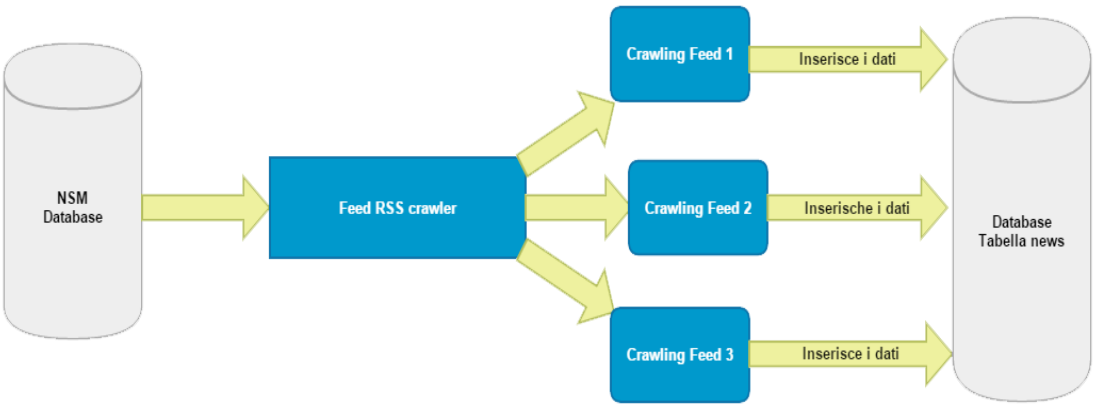


Figura 3.8: Mostra i passaggi per estrarre informazioni dal feed

3.10 Facebook Diffusion Crawler

Lo script appositamente creato (diffusion_fb.php) in PHP contiene tutti i comandi. Partendo dalla verifica dei parametri di connessione il crawler va ad analizzare le url di tutte le notizie prese dal Feed Rss crawler.

Esempio 3.9: Funzione per prendere i link scaricati dal Feed Rss crawler

```
function getLink(&$conn, $dateSession, $maxAge, $limit = 1) {
        $query = "SELECT link FROM news WHERE (TIMESTAMPDIFF(SECOND,
lastCrawlerDate, NOW()) >= $maxAge - 1) OR ((UNIX_TIMESTAMP('$dateSession')
% $maxAge) <= 30) ORDER BY RAND() LIMIT $limit";
        $result = mysqli_query($conn, $query);
        if(!$result) return false;
        $response = array();
        while(list($link) = mysqli_fetch_row($result)) {
            $response[] = $link;
        };
        return $response;
    };
```

Una volta presi gli url, li invia a Facebook attraverso l'apposita funzione creata FqlQuery(FQL) residente nel file common_fb.php

Esempio 3.9.1: Funzione per prendere i link delle news e inviarle a Facebook

```
function fqlQuery($links, $api, $secret) {
      $urls = "'".implode("','", $links)."'";
      $fql  = "SELECT url, click_count, comment_count, comments_fbid, commentsbox_count,
like_count, normalized_url, share_count, total_count FROM link_stat WHERE url IN ($urls)";
      $apifql
="https://api.facebook.com/method/fql.query?format=json&query=".urlencode($fql)."&access_token
=$api|$secret";
      $fb_json=file_get_contents($apifql);
      return json_decode($fb_json);
}
```

Una volta che Facebook ha ricevuto la chiamata, la analizza ed estrapola i dati da noi richiesti che successivamente vengono inseriti nel database di NSM(News on social media) nelle tabelle link_diffusion e link_history.
Terminato, nel database lo stato della notizia si aggiornerà cambiando da downloading a complete.

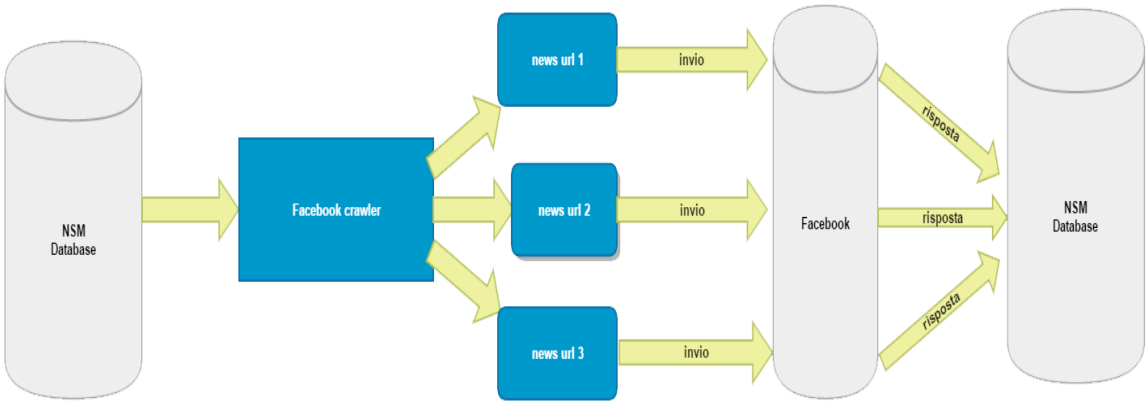


Figura 3.9: Mostra i passaggi per inviare l'url delle news a Facebook ed inserisce le risposte nel database.

3.11 Analisi

Allo stato attuale le analisi sono state fatte in tempo reale e permettono solo di creare particolari viste dei dati e questi sono solo degli approcci esplorativi rispetto ai dati raccolti.
Per la visione dei dati sono state create delle API (Application Programming Interface), script in php per l'estrazione di informazioni dal database.
Sono stati considerati campioni di dati da inizio aprile ad ora, quindi vi è possibile sia una visione giornaliera di questi che temporale.
In alcuni grafici, soprattutto quelli spline, possiamo vedere dei picchi verso l'alto o il basso e corrispondono ad eventi reali che possono essere accaduti in quel

determinato periodo(es: un grande incremento di condivisioni delle notizie con categoria “Scuola” dato l’introduzione di nuove regole per la scuola).

Nel progetto sono state sviluppate diverse API:

Nome File	Descrizione
getActiveOnlFace.php	Prende le notizie giornalistiche più attive su Facebook prendendo il campo "total_count", ovvero la somma dei count di facebook(commenti, like, condivisioni, ..) restituendolo con il nome della testata
getActiveOnlTestate.php	Quali sono le testate giornalistiche online più attive esaminando la tabella Link_diffusion sul database
getActiveOnlTestateHist.php	Prende le testate giornalistiche online più attive esaminano la tabella link_history sul database
getCategoriesDiffusion.php	Prende le categorie più attive esaminano la tabella link_diffusion sul database
getCategoriesDiffusionHist.php	Prende le categorie più attive esaminano la tabella link_history sul database per un range più ampio!
getIntDesSin.php	Verifica ed estrapola quali testate ricevono più interazioni tra quelle di destra, sinistra o centro.
getIntGenSpec	Verifica ed estrapola quali testate ricevono più interazioni tra quelle generalistiche e specialistiche.
getLinksDiffusionsHist.php	Prende tutti i link, titolo della news e descrizione dalla tabella news, con annesso
getNoDiff.php	Conta tutte le notizie che hanno i count a zero facendo una media con quelle che lo hanno
getNewsLifeHist.php	Prende i dati relativi alla notizia scelta(total_count, click count, comment count, share_count e like count) e fa vedere la sua storia nel tempo

4. User Interface

L'interfaccia grafica è stata realizzata con: HTML5, Css3, Bootstrap3, e Jquery, invece per la visualizzazione dei grafici è stata integrata la libreria di Highcharts/Highstock.

CSS3 è la nuova versione di CSS (Cascading Style Sheets) chiamato in italiano (Fogli di Stile a Cascata), è un linguaggio di programmazione web utilizzato per descrivere l'aspetto e la formattazione di un sito web al browser lato client.

È stato usato utilizzato per introdurre i media query [11] permettendo così la visualizzazione dell'interfaccia web pure sui dispositivi mobili.

CSS3 permette agli sviluppatori di creare pagine web ricche di contenuti con requisiti di codice relativamente leggeri. Ciò significa che gli effetti visivi sono più elaborate, pagine più leggere, migliore interfaccia utente, e caricamento più veloce delle pagine.

Il nuovo CSS3 è completo di elementi grafici come ombre, sfumature, effetti di bordo, layout multi-colonna e molto altro. Un esempio con una sua propietà @media per rendere l'interfaccia web accessibile a qualsiasi dispositivo.

```
@media (min-width: 200px) and (max-width: 480px) {
   select{
        width: 75%;
   }
    body {
        padding-top: 50px;
        width: 100%;
        height: 100%;
        font-family: 'Montserrat', sans-serif;
        font-size: 8px;
        margin-left: 10px;
   }


#grafo{
```

[11] Le media query sono filtri intuitivi applicabili agli stili CSS che aiutano a modificarli in base alle caratteristiche al dispositivo che esegue il rendering di contenuti, come ad esempio tipo, larghezza, altezza, orientamento e risoluzione del display.

```
        width: 55%;
        height: 300px;
    }
    #contenitore{
         padding-bottom: 350px;
    }
}
```

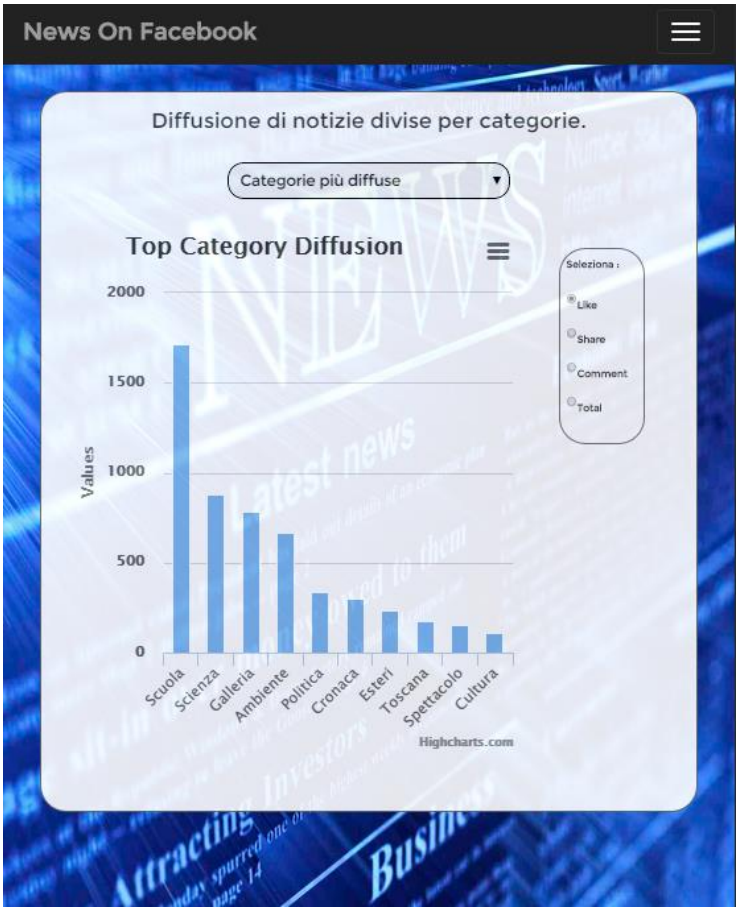


Figura 4: Una visualizzazione dell'intefaccia creata in versione mobile

5. Risultati

Data la grande quantità di dati ottenuti dall'analisi del periodo gennaio-maggio 2015, che rallenterebbero la visualizzazione dei risultati, abbiamo preso in esame un campione degli stessi, nel periodo 8 Gennaio 2015, 2 Marzo 2015.
Attraverso l'utilizzo della user interface visualizziamo una serie di risultati come:

- Testate più attive
- Categorie
- Analisi delle news
- Statistiche testate

Le **testate più attive** si compongono di tre sottocategorie:

1. *Testate più attive*: visualizza un grafico a barre riassuntivo dove per ciascuna testata giornalistica restituisce il numero di notizie in ordine decrescente.

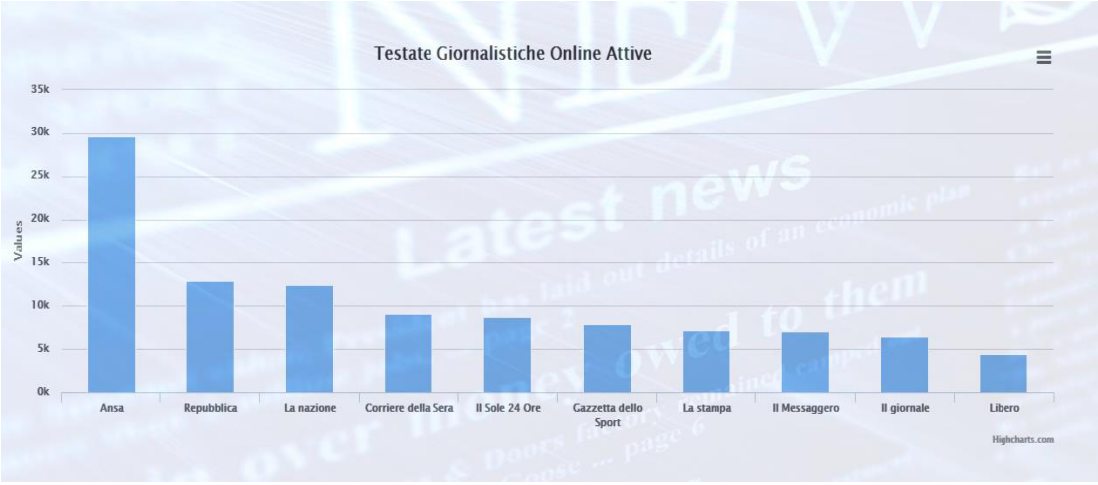


Figura 5: Una visualizzazione delle news

Il sito Ansa ha pubblicato, nel periodo di tempo in esame, il maggior numero di notizie rispetto alle testate giornalistiche prese in esame

2. *Testate più diffuse su facebook*: visualizza un grafico a barre riassuntivo dove considera le notizie più attive su Facebook prendendo il campo "total_count", ovvero la somma dei count di facebook(commenti, like, condivisioni, ..) per ogni singola testata in ordine decrescente. È possibile inoltre selezionare i risultati per “like”, “share”, “comment”, “total”.

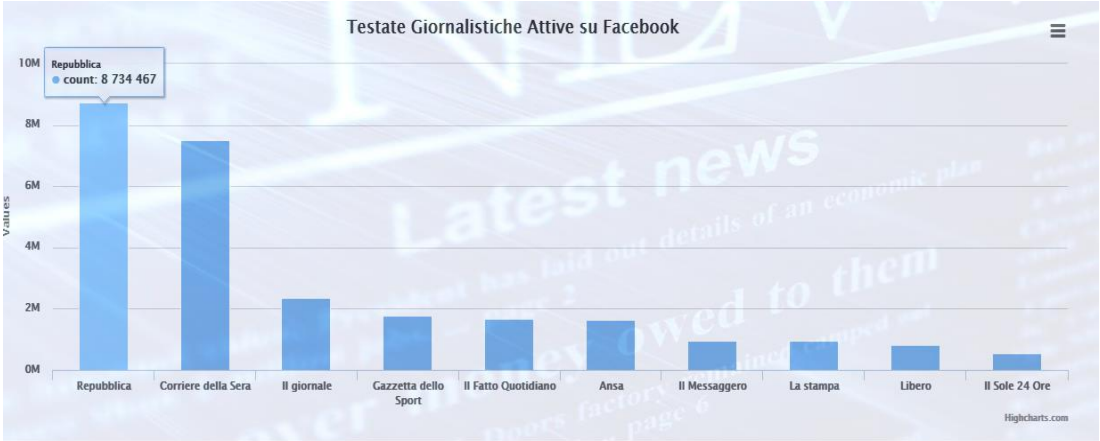


Figura 5.1: Una visualizzazione delle news

Il grafico evidenzia le testate giornalistiche diffuse attraverso facebook con più “like” in ordine decrescente: Repubblica risulta essere la più “piaciuta”.

3. *History della diffusone di notizie*: un grafico spline dove per ciascuna testata giornalistica raffronta la somma dei like o share o comment o total verso il numero delle notizie di un determinato periodo di tempo. È possibile inoltre selezionare i risultati per “like”, “share”, “comment”, “total”.

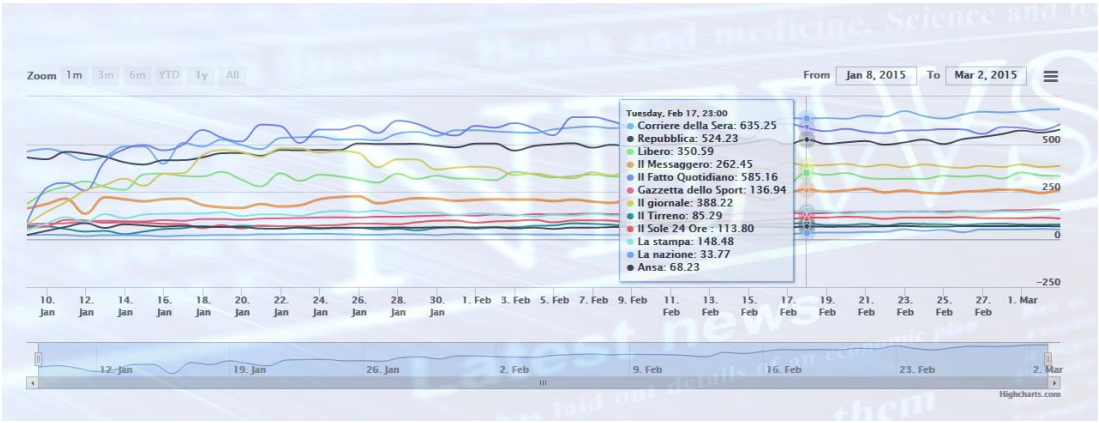


Figura 5.2: Una visualizzazione delle news

Il grafico evidenzia Il corriere della sera come la testata giornalistica più attiva attraverso i like.

Le **Categorie** si compongono di tre sottocategorie:

1. *Categorie più diffuse:* visualizza un grafico a barre riassuntivo dove per singolo argomento(scuola, ambiente, politica, scienza, ..) raffronta la somma dei like o share o comment o total verso il numero delle notizie giornaliere.

Figura 5.3: Una visualizzazione delle categorie

Il grafico evidenzia che la scuola è la categoria più diffusa.

2. *History delle categorie:* visualizza un grafico spline dove per singolo argomento(scuola, ambiente, politica, scienza, ..) raffronta la somma dei like o share o comment o total verso il numero delle notizie di un determinato periodo di tempo.

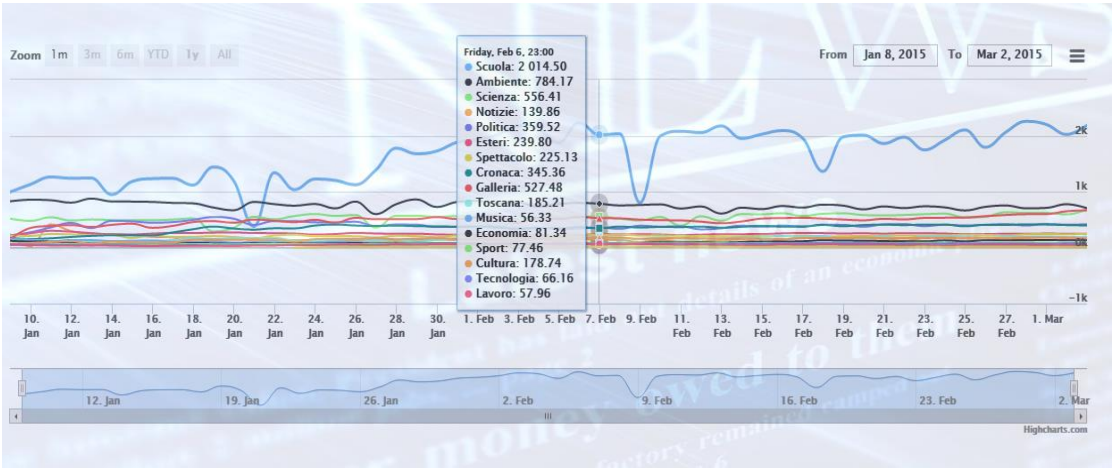


Figura 5.4: Una visualizzazione storica delle categorie

Il grafico evidenzia che la categoria di notizie più diffusa è la scuola seguita da ambiente e scienza.

3. *Categorie per news:* visualizza un grafico a barre riassuntivo dove per ciascuna categoria restituisce il numero di notizie in ordine decrescente.

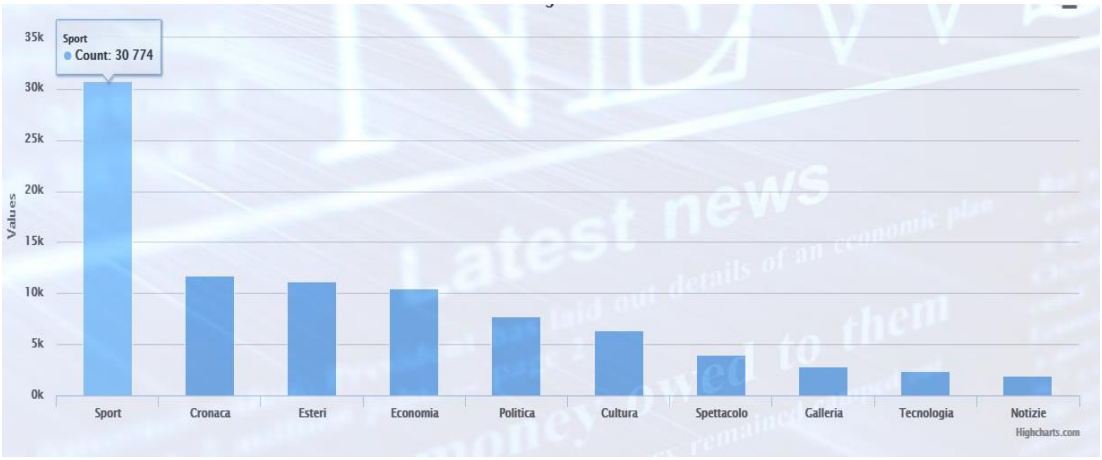


Figura 5.5: Una visualizzazione delle categorie

Il grafico evidenzia il maggior numero di notizie pubblicate per la categoria dello sport seguito da cronaca e esteri.

Le **Analisi delle news** su Facebook si basano sulla diffusione o non diffusione di una determinata notizia da parte degli utenti.

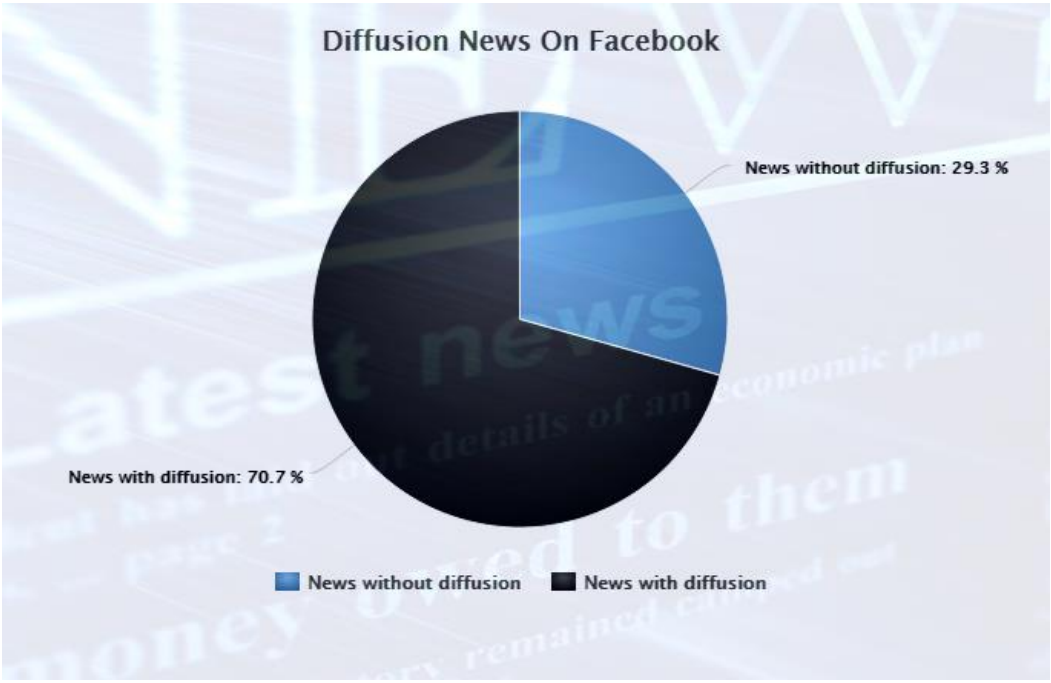
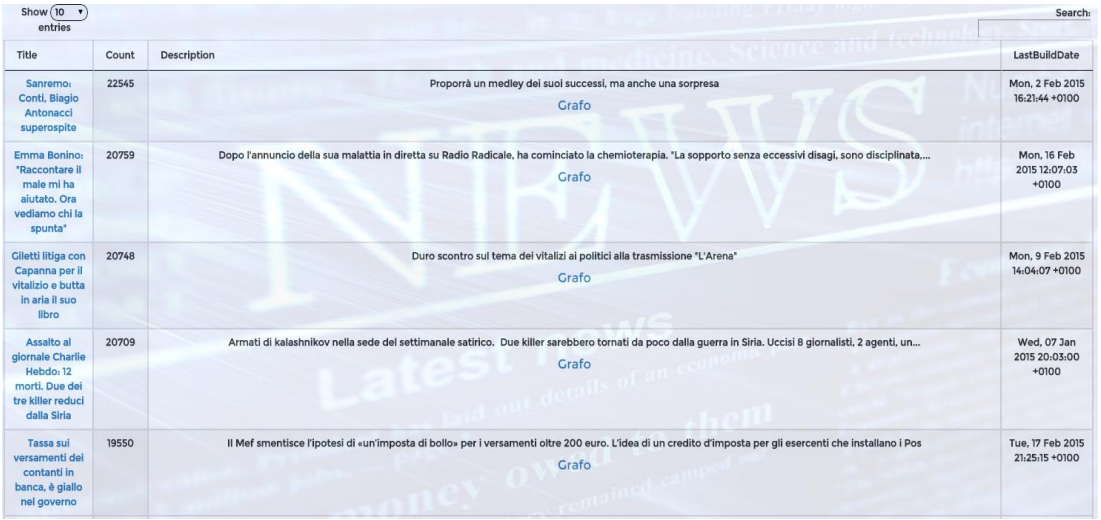


Figura 5.6: Una visualizzazione delle news

Viene generata una tabella contenente i dati: titolo notizia, url della notizia, contatore, descrizione e la data di pubblicazione notizia.
Successivamente è possibile visualizzare un grafico con i relativi count (like, share, comment, total) della notizia desiderata.


Show 10 entries

Search:

Title	Count	Description	LastBuildDate
Sanremo: Conti, Biagio Antonacci superospite	22545	Proporrà un medley dei suoi successi, ma anche una sorpresa Grafo	Mon, 2 Feb 2015 16:21:44 +0100
Emma Bonino: "Raccontare il male mi ha aiutato. Ora vediamo chi la spunta"	20759	Dopo l'annuncio della sua malattia in diretta su Radio Radicale, ha cominciato la chemioterapia. "La sopporto senza eccessivi disagi, sono disciplinata,... Grafo	Mon, 16 Feb 2015 12:07:03 +0100
Giletti litiga con Capanna per il vitalizio e butta in aria il suo libro	20748	Duro scontro sul tema dei vitalizi ai politici alla trasmissione "L'Arena" Grafo	Mon, 9 Feb 2015 14:04:07 +0100
Assalto al giornale Charlie Hebdo: 12 morti. Due dei tre killer reduci dalla Siria	20709	Armati di kalashnikov nella sede del settimanale satirico. Due killer sarebbero tornati da poco dalla guerra in Siria. Uccisi 8 giornalisti, 2 agenti, un... Grafo	Wed, 07 Jan 2015 20:03:00 +0100
Tassa sui versamenti dei contanti in banca, è giallo nel governo	19550	Il Mef smentisce l'ipotesi di «un'imposta di bollo» per i versamenti oltre 200 euro. L'idea di un credito d'imposta per gli esercenti che installano i Pos Grafo	Tue, 17 Feb 2015 21:25:15 +0100

Figura 5.7: Una visualizzazione per singola news

Per la tabella è possibile scegliere quante notizie visualizzare attraverso il menù a tendina in alto a sinistra e cercare un determinata notizia con il search in alto a destra.

A fine tabella, in basso a destra, sarà visibile la paginazione.

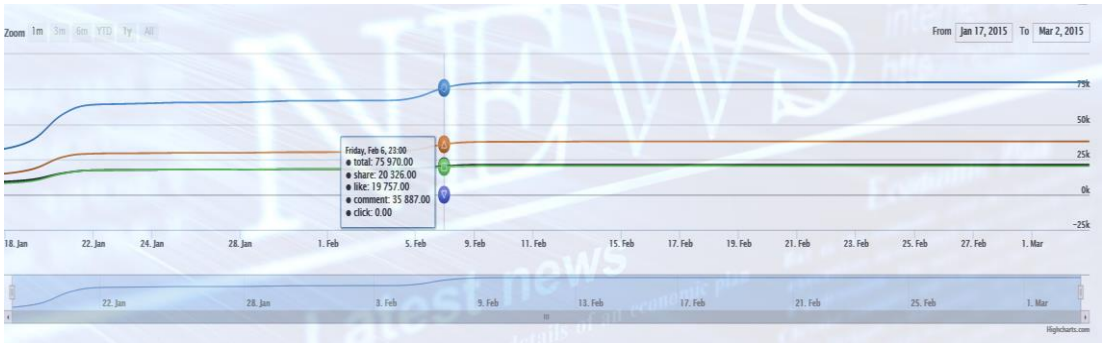


Figura 5.8: Una visualizzazione storica per singola news

Inoltre è possibile vedere la history della notizia interessata attraverso un altro grafico, che considera tutte le statistiche da quando è stata pubblicata la notizia.

Le **Statistiche delle testate** rappresentano la distribuzione statistica delle testate giornalistiche visualizzate in generalistiche vs specialistiche e per diverso orientamento politico.
È possibile inoltre selezionare i risultati per “like”, “share”, “comment”, “total”.

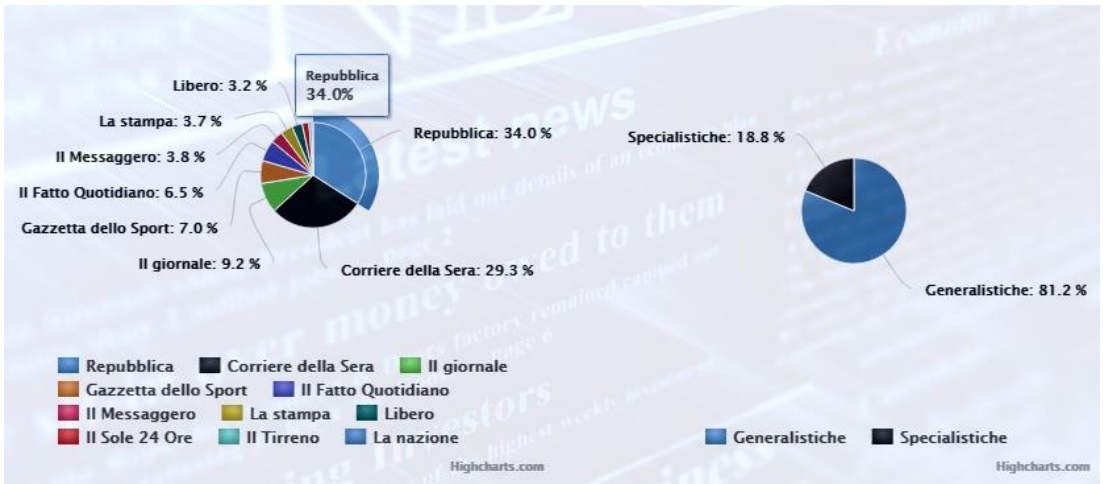


Figura 5.9: Una visualizzazione per testate giornalistiche

In questo risultato si può vedere la testata giornalistica più “Piaciuta”, Repubblica e il confronto tra testate specialistiche e generalistiche.

6. Conclusioni

Questo strumento di analisi permette di analizzare e organizzare in modo logico i dati statistici che vengono generati dall'utilizzo delle notizie diffuse in rete dalle testate giornalistiche.

Il lavoro realizzato si basa sulle API offerte da Facebook. Quindi un punto di debolezza della piattaforma deriva dall'affidabilità dei dati di ricavati dalle API.

Il sistema, nonostante sia giovane, riesce a dare un'immagine della realtà e può essere utile per studiare in maniera retrospettiva le tendenze della popolazione.

È inoltre uno strumento che può essere ulteriormente implementato con altre testate giornalistiche e che quindi può aumentare il numero di notizie da monitorare.

Ciò nonostante il sistema deve essere migliorato per ottenere una maggiore scalabilità ed occorre identificare un criterio di "stop" quando non raccoglie più informazioni di una notizia in modo da evitare la saturazione del database(crash).

Lo strumento si può estendere ad altri social network anche se questo lavoro si è concentrato esclusivamente su Facebook perché con le API l'acquisizione dei dati è più semplice oltre al fatto che Facebook essendo il social network più diffuso in Italia è anche fonte di campioni eterogenei.

7. Ringraziamenti

Desidero ricordare e ringraziare tutti coloro che mi hanno sostenuto ed aiutato durante il mio corso di studi, soprattutto quelle persone che ho sempre avuto vicino: la mia ragazza Arianna, il mio affezionatissimo “team” di amici e compagni, in particolar modo, Nicholas, Gianluca, Giulio e gli amici con cui sono nato e cresciuto e che ancora oggi mi sopportano: Leo, Sacha, Luca.

Un ringraziamento particolare va a mia madre e mio padre che hanno avuto la pazienza di sopportarmi e supportarmi, la loro presenza è stata fondamentale e io gli sono molto grato, a cui dedico questo lavoro per l’amore e la forza che hanno saputo darmi ogni giorno.

Insieme a loro ringrazio mio fratello Stefano e tutti i familiari, soprattutto la mia nonna Fiorella.

Un ringraziamento particolare va anche all’istituto di Informatica e Telematica del CNR di Pisa per avermi dato la possibilità di lavorare a questo progetto.

Ringrazio il professore Andrea Marchetti, Mirko Tavosanis e Davide Gazzè, per i loro preziosi consigli sull’implementazione e il tempo dedicatomi.

Bibliografia

[1] Tesconi, Maurizio, Davide Gazzé, and Angelica Lo Duca. "SocialTrends: a web application for monitoring and visualizing users in social media." Social Informatics. Springer Berlin Heidelberg, 2012. 535-538.

Sitografia

- Social Network,Wikipedia: http://it.wikipedia.org/wiki/Servizio_di_rete_sociale (visitato 11Maggio 2015).
- Feed Rss, Wikipedia: http://it.wikipedia.org/wiki/RSS (visitato 11 Maggio 2015).
- Repubblica, testata giornalistica online: http://www.repubblica.it/ (visitato 11 Maggio 2015).
- Ansa, agenzia notize online: http://www.ansa.it/sito/static/ansa_rss.html (visitato 11 Maggio 2015).
- Corriere della sera, testata giornalistica online: http://www.corriere.it/ (visitato 11 Maggio 2015).
- Il Fatto Quotidiano, testata giornalistica online: http://www.ilfattoquotidiano.it/feed/ (visitato 11 Maggio 2015).
- Il Tirreno, testata giornalistica online: http://iltirreno.gelocal.it/utility/2007/06/27/news/feed-rss-1.1710820 (visitato 11 Maggio 2015).
- La nazione, testata giornalistica online: http://qn.quotidiano.net/varie/2012/03/30/689565-Mappa-sito-feed-RSS.shtml (visitato 11 Maggio 2015).

- La Stampa, testata giornalistica online: http://www.lastampa.it/rss (visitato 11 Maggio 2015).
- Il messaggero, testata giornalistica online: http://www.ilmessaggero.it/rss.php (visitato 11 Maggio 2015).
- Il giornale, testata giornalistica online: http://www.ilgiornale.it/tag/feed-rss.html (visitato 11 Maggio 2015).
- Libero, testata giornalistica online: http://www.liberoquotidiano.it/feed.jsp (visitato 11 Maggio 2015).
- La Gazzetta dello sport, testata giornalistica online: http://www.gazzetta.it/rss/ (visitato 11 Maggio 2015).
- Il Sole 24 ore, testata giornalistica online: http://www.ilsole24ore.com/st/extra/rss.shtml (visitato 11 Maggio 2015).
- Trackur, Social media monitor: http://www.trackur.com (visitato 11 Maggio 2015).
- SocialMention, Social media monitor: http://socialmention.com/ (visitato 11 Maggio 2015).
- Cyberalert, Social media monitor: http://www.cyberalert.com (visitato 11 Maggio 2015).
- Hootsuite, Social media monitor: https://hootsuite.com/ (visitato 11 Maggio 2015).
- Morover Newdesh, Social media monitor: http://www.moreover.com/newsdesk/media-monitoring/ (visitato 11 Maggio 2015).
- Social Trends, Social media monitor: http://www.social-trends.it/home (visitato 11 Maggio 2015).

- KPI6, Social media monitor: https://kpi6.com/ (visitato 11 Maggio 2015).
- Flipboard, Social media magazine: https://flipboard.com/ (visitato 11 Maggio 2015).
- Bootstrap, Libreria: http://getbootstrap.com/ (visitato 11 Maggio 2015).
- Jquery, Libreria: https://jquery.com/ (visitato 11 Maggio 2015).
- Highcharts/Highstock, Libreria: http://www.highcharts.com/ (visitato 11 Maggio 2015).
- Simplexml, Libreria: http://php.net/manual/en/book.simplexml.php (visitato 11 Maggio 2015).
- FacebookFQL: https://developers.facebook.com/docs/reference/fql/ (visitato 11 Maggio 2015).