POLITECNICO DI MILANO

Scuola di Ingegneria Industriale e dell'Informazione

Corso di Laurea Magistrale in Ingegneria Matematica

Analisi dei social network: Sentiment e Opinion Analysis applicate agli atenei italiani

Relatore:

Prof. Simone Vantini

Correlatore:

Prof.ssa Marika Arena

Candidato:

Riccardo Mignogna

805462

Anno Accademico 2014/2015

Abstract

L'esplosione dei dati provenienti dai *blogs*, *forum online*, *social network*, permette, a chi in possesso dei giusti strumenti di analisi, di individuare il *sentiment* e le preferenze degli utenti sui vari eventi, prodotti o servizi in genere.

Lo scopo di questa tesi è l'applicazione delle tecniche di *Sentiment Analysis* e *Opinion Analysis* per i messaggi relativi ai diversi atenei italiani provenienti dai due *social network* principali: Facebook e Twitter.

Lo studio ha previsto una prima fase di *hand-coding* dei dataset relativi al *case study* in esame. Successivamente si è proceduto ad analisi dei parametri dei metodi matematici utilizzati e il loro relativo *tuning*.

Vengono successivamente esaminati in dettaglio i risultati conseguiti, ricercando inoltre un modello matematico in grado di cogliere l'evoluzione dei sentimenti espressi dagli studenti sui *social network*.

In conclusione vengono evidenziate tutte le criticità incontrate, i problemi ancora aperti e le possibili migliorie per conseguire risultati più precisi e accurati.

Indice

Indice delle figure **iii**

Indice delle tabelle **v**

Introduzione **1**

1 I Dati **3**

1.1 I post su Facebook 3

1.2 I tweet su Twitter 5

2 Analisi Testuale **9**

2.1 Classificazione individuale e aggregata 9

2.2 Preprocessamento del testo 10

2.3 Il metodo di Hopkins e King 13

2.3.1 Selezione delle categorie 15

2.3.2 Hand-Coding 16

2.3.3 Ipotesi e loro verifica 17

2.3.4 Bagging 18

2.3.5 Vantaggi e svantaggi 19

3 *Case Study*: Gli atenei italiani **21**

3.1 Analisi 23

3.1.1 Categorie 23

3.1.2 Tagging manuale 25

3.1.3 Misura di accuratezza 27

3.1.4 Il pacchetto *ReadMe* 27

3.2 Twitter 29

3.2.1 Il training set e l'hand-coding 30

3.2.2 Elaborazione dei *tweet* 31

3.2.3 Sistema Universitario 32

3.2.4 Confronto tra gli atenei 35
3.2.5 Singola università: Politecnico di Milano 40
3.3 Facebook 41
3.3.1 Il training set e l'hand-coding 42
3.3.2 Elaborazione dei *post* 43
3.3.3 Sistema Universitario 44
3.3.4 Confronto tra gli atenei 47
3.3.5 Singola università: Politecnico di Milano 53
3.4 Regressione Logistica 55
3.4.1 Regressori 56

4 Conclusioni **61**

Appendice A **65**

Appendice B **67**

Bibliografia **69**

Indice delle figure

1 L'Italia nei due principali social network. Immagine tratta da vincos.it. 1

2.1 Processo di *Classify and Count* 9

3.1 *Sentiment and Opinion Analysis*: dai testi ai *report* 21
3.2 I dati dai due social network 22
3.3 Distribuzione dei *tweet* per ateneo e per istituzionalità . . . 29
3.4 Numero *tweet* per le categorie di *sentiment analysis* nel *training set* 30
3.5 Numero *tweet* per le categorie di *opinion analysis* nel *training set* 31
3.6 Distribuzione dei *tweet* per le categorie di *sentiment analysis* per l'intero sistema universitario 33
3.7 Confronto *tweet* con giudizio per l'intero sistema universitario per la *sentiment analysis* 33
3.8 Confronto *tweet* con giudizio per l'intero sistema universitario per l'*opinion analysis* 34
3.9 Ranking *tweet* con giudizio tra i diversi atenei per la *sentiment analysis* 35
3.10 Benchmarking *tweet* istituzionali con giudizio tra i diversi atenei per la *sentiment analysis* 36
3.11 Benchmarking *tweet* non istituzionali con giudizio tra i diversi atenei per la *sentiment analysis* 37
3.12 Distribuzione dei *tweet* per il Politecnico di Milano 40
3.13 Analisi temporale dei *tweet* per il Politecnico di Milano . . . 40
3.14 Distribuzione dei *post* per ateneo e per istituzionalità 41
3.15 Numero *post* per le categorie di *sentiment analysis* nel *training set* 42

3.16 Numero *post* per le categorie di *opinion analysis* nel *training set* 43
3.17 Distribuzione dei *post* per le categorie di *sentiment analysis* per l'intero sistema universitario 44
3.18 Confronto *post* con giudizio per l'intero sistema universitario per la *sentiment analysis* 45
3.19 Confronto *post* con giudizio per l'intero sistema universitario per l'*opinion analysis* 46
3.20 Ranking *post* con giudizio tra i diversi atenei per la *sentiment analysis* 47
3.21 Benchmarking *post* istituzionali con giudizio tra i diversi atenei per la *sentiment analysis* 48
3.22 Benchmarking *post* non istituzionali con giudizio tra i diversi atenei per la *sentiment analysis* 49
3.23 Distribuzione dei *post* per il Politecnico di Milano 53
3.24 Analisi temporale dei *tweet* del 2014 per il Politecnico di Milano 54
3.25 Analisi temporale dei *tweet* del 2015 per il Politecnico di Milano 54

Indice delle tabelle

1.1 Esempi di post da Facebook 5
1.2 Esempi di tweet da Twitter 6
1.3 Suddivisione degli atenei in esame nei due social network e secondo il criterio di istituzionalità 7

2.1 Esempio di *preprocessing* di un testo 11
2.2 Esempio di una matrice di stemming 12
2.3 Categorie solite per la *Sentiment Analysis* 15

3.1 Categorie *Sentiment Analysis* del *Case Study* 23
3.2 Categorie *Opinion Analysis* del *Case Study* 24
3.3 Grado di accordo tra i diversi *Coders* 25
3.4 Analisi descrittiva dataset Twitter 29
3.5 *Opinion Analysis* per i *tweet* istituzionali 38
3.6 *Opinion Analysis* per i *tweet* non istituzionali 39
3.7 Analisi descrittiva dataset Facebook 41
3.8 *Opinion Analysis* per i *post* istituzionali 51
3.9 *Opinion Analysis* per i *post* non istituzionali 52
3.10 Valori dei vari regressori per ogni ateneo nello studio di Facebook 57
3.11 Valori dei vari regressori per ogni ateneo nello studio di Twitter 58

4.1 Risultati regressioni logistiche nello studio di Facebook: VSOth altri post con giudizio, VSAll contro tutti gli altri post . . . 65
4.2 Risultati regressioni logistiche nello studio di Twitter: VSOth altri post con giudizio, VSAll contro tutti gli altri post . . . 66

Introduzione

Le università italiane sui social network

Con il continuo aumento dell'utilizzo della rete, i social network sono diventati una presenza costante nella vita di tutti i giorni. Sempre più italiani usano web e social network per informarsi e condividere le notizie. Secondo un rapporto del Censis, il 71% della popolazione italiana è attiva su internet, inoltre, è iscritto a Facebook circa il 50% dell'intera popolazione, percentuale che cresce siano al 77% considerando i soli giovani under 30 (categoria dove ricade la maggior parte degli studenti universitari). Mentre per quanto riguarda Twitter circa il 15% degli italiani possiede un account e il 6% è un utente attivo (almeno un'attività nel corso del mese).

Tramite i social network gli utenti possono, oltre a riportare come in un diario i momenti della loro giornata, esprire pareri e giudizi su prodotti e/o servizi utilizzati. Lo studio Oppici, Martin, Morando, Basso & Futia (2014) ha rilevato che l'80% delle università ha almeno un account Facebook, mentre tale percentuale si attesta al 76% su Twitter. Gli atenei utilizzano i social network, principalmente per la promozione e documen-

(a) Facebook in Italia **(b)** Twitter in Italia

Figura 1: L'Italia nei due principali social network. Immagine tratta da vincos.it.

tazione degli eventi interni, la pubblicazione di promemoria per scadenze burocratiche, fino alle attività di segreteria e sportello d'informazione per gli studenti. Di conseguenza, come per un'azienda produttrice o un'erogatrice di servizi diventa fondamentale usare le informazioni che arrivano dalla rete, anche per le università è essenziale l'analisi di questi dati. Infatti, attraverso l'analisi dei testi (post nel caso di Facebook o tweet per Twitter) è possibile ottenere informazioni sulle possibili azioni effettuabili dall'ateneo per migliorare il proprio rapporto con gli studenti.

Sentiment e Opinion Analysis

Proprio per via della crescente diffusione dei social media, l'analisi testuale ha subito nel corso degli ultimi anni un'attenzione sempre maggiore. La *Sentiment Analysis* si riferisce al riconoscimento, attraverso strumenti matematici e statistici, del parere o giudizio espresso dagli utenti. I più classici esempi di *sentiment* sono *positivo*, *negativo* o *neutro* ma nulla vieta di utilizzare una gamma di pareri diversi (*contento*,*triste*,...).
L'*Opinion Analysis* è un'indagine sulla natura di un testo, ovvero mira a riconoscere le motivazioni alla base di un particolare *sentiment*. Anche prima dell'esplosione dei social media, strumenti come sondaggi e/o servizi al cliente, permettevano di indagare il *sentiment*. L'analisi dei social media ha però diversi vantaggi come: celerità nell'ottenimento dei dati, costi molto ridotti, campione molto ampio ed eterogeneo e non ultimi la spontaneità e sincerità dei pareri espressi. Di contro, raramente, si hanno informazioni sul campione indagato, come l'età, il sesso, occupazione che possono risultare rilevanti.

Lo scopo di questo elaborato sarà quello di analizzare il *sentiment* e la natura dello stesso (*opinion*) degli studenti universitari nei confronti del proprio ateneo di appartenenza. A tale scopo si utilizzerà il metodo introdotto da Hopkins & King (2010). Occorre però sottolineare che non esiste la tecnica di analisi testuale, ma ogni metodo ha performance diverse a seconda dei dati a disposizione e degli obiettivi preposti.

Capitolo 1

I Dati

Il caso di studio in oggetto prevede l'indagine dei giudizi, e la natura degli stessi, degli studenti universitari di ogni singolo ateneo in esame e del sistema universitario complessivo. I social network analizzati sono Facebook e Twitter, scelti per la massiccia presenta degli atenei al loro interno. La tabella 1.3 a pagina 7 mostra gli atenei presi in esame e le numerosità in gioco nei due diversi social.

1.1 I post su Facebook

Per raccogliere i dati sui diversi atenei provenienti da Facebook (FB), si è dovuto prima procedere ad una fase di mappatura delle diverse presenze sul social network. Su Facebook sono presenti tre tipologie principali di accesso:

- **Profilo privato:** La singola persona collegata a FB, con un diario personale e con la possibilità di commentare in pagine e gruppi aperti.

  Esempio: Pagina personale di un singolo studente del Politecnico di Milano;

- **Pagina:** Dedicata ad una singola persona o a istituzione (come le università), ma pubblica a qualsiasi utente di FB. Tipicamente utilizzata dai personaggi dello spettacolo, dello sport e da enti pubblici o privati.

  Esempio: La pagina del Politecnico di Milano in cui vengono pubblicati avvisi ed eventi;

- **Gruppo:** Spazio su FB riservato ad un insieme di utenti. In questo caso è possibili definire diversi livelli di privacy, tra cui se il gruppo sarà aperto (visibile e accessibile da chiunque) o chiuso (visibile ma non accessibile). Nel caso dei gruppi chiusi non è possibile utilizzare i dati (post) al loro interno.

  Esempio: Il gruppo (aperto) di Ingegneria Matematica;

In questo studio si utilizzeranno come sorgenti dati, i post pubblici e quindi fruibili di FB presenti sulle pagine e gruppi aperti relativi agli atenei in esame. I post sono stati raccolti tramite Netvizz presente su FB, che permette di censire e collezionare tutti i post pubblici delle pagine e gruppi aperti richiesti. Inoltre si è considerata una suddivisione dei post tra istituzionali o non istituzionali: i post istituzionali sono scritti e condivisi da soggetti autorizzati dall'ateneo. Si è realizzato dunque un foglio di calcolo contente 97595 post, a cui sono associate le seguenti informazioni:

- **Ateneo ID:** Codice identificativo dell'ateneo;
- **Ateneo:** Nome ateneo;
- **P/GA:** Se la sorgente del post è una Pagina o un Gruppo Aperto;
- **I/NI:** Se il post è considerato istituzionale o non istituzionale;
- **Type:** Se il post è uno status, immagine o un link. Verranno analizzati i soli post di status, ovvero quei post che presentano un testo;
- **Post Message:** Il testo presente nel post;
- **Post Published:** Il giorno di pubblicazione del post;

Nella tabella 1.1 nella pagina successiva si riportano qualche esempio di post scaricati da FB.

L'arco temporale di pubblicazione dei post raccolti varia dal primo Gennaio 2014 al 30 Giugno 2015. Le numerosità dei post negli atenei e nella suddivisione tra istituzionali e non, sono riportati nella tabella 1.3 a pagina 7.

Ateneo	P/GA	I/NI	Post Message	Post Published
Ca Foscari	P	I	Salve, un'informazione: il bando Erasmus è uscito? Ed eventualmente dove posso trovarlo?	11/02/2014
Politecnico di Milano	GA	NI	Ciao! sto cercando un Bramanti - Pagani - Salsa analisi II.	17/05/2015
Bicocca	GA	NI	Qualcuno sa dirmi come è psicologia dello sviluppo?	26/05/2015

Tabella 1.1: Esempi di post da Facebook

1.2 I tweet su Twitter

A differenza di Facebook, su Twitter ogni utente può condividere solo brevi messaggi di testo, fino ad un massimo di 140 caratteri, chiamati *tweet*. Sono degli aggiornamenti visibili nella pagina profilo dell'utente e che vengono inoltrati a tutti coloro che si sono registrati per riceverli, ossia ai *follower* di questo utente. Su Twitter vige la regola del "silenzio assenso", ovvero un utente può "seguire" un altro senza bisogno della sua autorizzazione, a meno che quest'ultimo non abbia deciso di rendere privato il proprio account. Nonostante la base utenti di Twitter sia solo una frazione di quella di Facebook, questo social media, proprio per le sue caratteristiche di apertura e orizzontalità, sta diventando una fonte di notizie in tempo reale estremamente influente sia per l'intera rete che per i media tradizionali. Anche in questo caso si sono suddivisi in tweet tra istituzionali o non istituzionali,a seconda se provenienti o meno da utenti autorizzati dall'ateneo. Si è realizzato dunque un ulteriore foglio di calcolo contente 56394 tweet, contenenti le seguenti informazioni:

- **User:** Nome ateneo;
- **I/NI:** Se il tweet proviene è considerato istituzionale o non istituzionale;
- **Text:** Il testo presente nel tweet;
- **Date:** Il giorno di pubblicazione del tweet;

Nella tabella 1.2 si riportano qualche esempio di tweet ottenuti da Twitter.

User	**I/NI**	**Text**	**Date**
Milano Statale	NI	La gente che non ha ancora capito dopo due anni che le mail dei prof sono: nome.cognome@unimi.it	05/03/2015
Scuola Normale	I	#Matematica matematici e Grande Guerra Un convegno in #Normale ora in streaming.	23/05/2015
Politecnico di Torino	NI	Come prevenire lo sversamento di petrolio negli oceani? L'idea degli studenti del Politecnico di Torino.	12/06/2015

Tabella 1.2: Esempi di tweet da Twitter

In questo caso l'arco temporale in esame dei tweet scaricati varia dal primo Gennaio 2015 al 30 Giugno 2015. Le numerosità dei tweet negli atenei e nella suddivisione tra istituzionali e non, sono riportati nella tabella 1.3 a fronte

Ateneo	FB Totale	FB Istit.	FB Non Istit.	TW Totale	TW Istit.	TW Non Istit.
Politecnico di Milano	6457	664	5793	9577	456	9121
Bicocca	8981	122	8859	4961	809	4152
Bologna	4930	272	4658	9374	2257	7117
Brescia	2747	15	2732	1081	190	891
Ca Foscari	6736	1124	5612	2656	977	1679
Ferrara	1326	52	1274	914	0	914
Insubria	2301	55	2246	583	0	583
IUAV	1267	197	1070	999	135	864
Milano Statale	10854	335	10519	7447	712	6735
Parma	6785	141	6644	1721	58	1663
Pavia	3699	275	3424	2960	699	2261
Piemonte Orientale	790	12	778	574	331	243
Politecnico di Bari	3990	175	3815	710	37	673
Politecnico di Torino	4211	55	4156	2796	222	2574
Salento	7467	227	7240	601	101	500
Sassari	1826	224	1602	684	76	608
Torino	7658	878	6780	0	0	0
Trento	7671	79	7592	1153	0	1153
Udine	2690	464	2226	2832	671	2161
Verona	5052	76	4976	1424	480	944
Sant'Anna	70	46	24	952	483	469
SISSA	70	6	64	43	42	1
SNS	17	14	3	2352	595	1757

Tabella 1.3: Suddivisione degli atenei in esame nei due social network e secondo il criterio di istituzionalità

Capitolo 2

Analisi Testuale

2.1 Classificazione individuale e aggregata

Quando l'obiettivo dello studio è conoscere la classe di appartenenza di ogni singolo testo si procede con una classificazione individuale, qualora invece si stiano analizzando un numero elevato di testi e si vogliono conoscere le proporzioni nelle varie classi si utilizza una classificazione aggregata. Evidentemente è possibile passare da una classificazione individuale ad una aggregata tramite un processo di *Classify and Count* come mostrato in figura 2.1. Il problema nell'utilizzo di un processo di questo tipo è che operando una classificazione individuale e poi contando il numero di testi assegnati a ciascuna categoria si amplificano gli errori di misclassificazione rispetto alla classificazione aggregata.

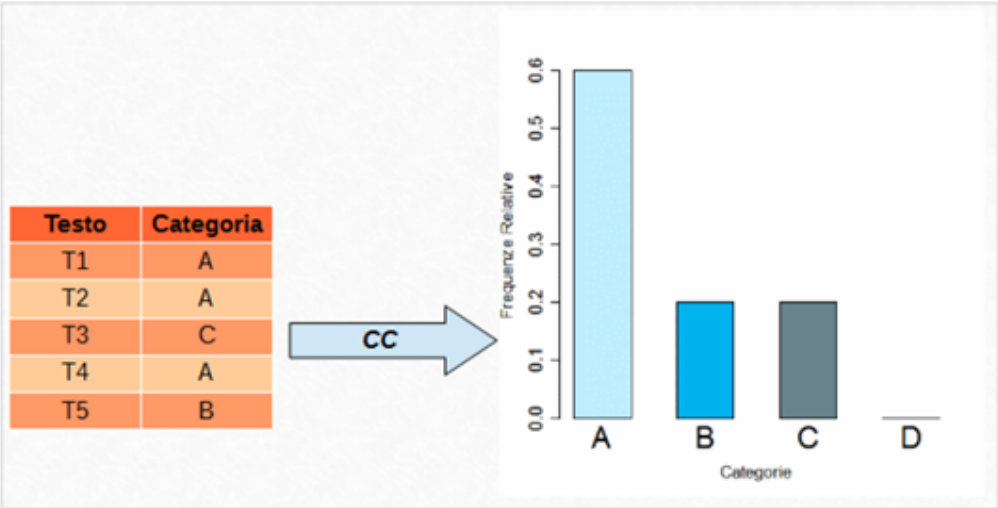


Figura 2.1: Processo di *Classify and Count*

2.2 Preprocessamento del testo

Il processo fondamentale alla base dell'analisi testuale, è la trasformazione del testo in un dato quantitativo tale da poter essere gestito da metodi automatici. In un testo sono contenuti molti termini o simboli che non portano significatività semantica e possono quindi essere filtrati attraverso una fase preliminare detta *preprocessing*. Ovviamente queste procedure causano una perdita di informazioni, come ad esempio l'ordine con cui le parole compaiono nel testo. Ciò nonostante una struttura troppo elaborata del metodo linguistico, inficerebbe la bontà dei metodi di analisi del testo. Le trasformazioni eseguibili su un testo sono innumerevoli, nel seguito si riporta una lista di quelle usate nel nostro caso.

- **Rimozione *stopwords*:** Molte parole all'interno di un testo non aggiungono nessuna informazione significativa per la comprensione del *sentiment*. Queste parole, dette *stopwords*, sono in genere preposizioni, articoli o verbi ausiliari. Individuare una lista di possibili *stopwords* non è semplice, quelle utilizzate nel nostro caso sono riportate in Appendice B.

- **Rimozione punteggiatura:** Nella comprensione di un testo, la punteggiatura risulta fondamentale, ma nell'utilizzo di un metodo statistico viene generalmente soppressa.

- **Riduzione in minuscolo:** Normalmente la modalità di scrittura (maiuscolo o minuscolo) non comporta alcuna informazione perciò si trasforma il testo tutto in minuscolo per utilizzare una rappresentazione univoca (anche se in alcuni contesti online come forum o blog, la scrittura in maiuscolo delle parole, indica un senso di rabbia o disapprovazione).

- **Rimozione spazi bianchi:** Vengono rimossi gli spazi bianchi in eccesso. Se tra due parole sono presenti più caratteri rappresentativi dello spazio bianco, vengono ridotti all'unità.

- **Stemming:** Nell'esprimere un concetto, in ogni lingua utilizzata, si adoperano diverse parole che presentano però la stessa radice. Ad esempio le parole *bello, bella, belli* possono essere sostituite della radice *bell* (che indichiamo con *stem*). Così facendo si ha una notevo-

le riduzione del linguaggio. L'algoritmo di *stemming* comunemente utilizzato è quello di Porter (1980).

Nella tabella 2.1 illustriamo un esempio di processamento del testo:

- *Il Politecnico parlerà inglese, un passo importante per il paese!*

Operazione	**Testo risultante**
Riduzione in minuscolo	il politecnico parlerà inglese, un passo importante per il paese!
Rimozione stopwords	__ politecnico parlerà inglese, __ passo importante ___ __ paese!
Rimozione punteggiatura	__ politecnico parlerà inglese_ __ passo importante ___ __ paese_
Stemming	__ politecn parl ingles_ __ pass import ___ __ paes_
Rimozione spazi bianchi	politecn parl ingles pass import paes

Tabella 2.1: Esempio di *preprocessing* di un testo

L'ultima fase di elaborazione del testo, è quella che ci permetterà di trattarlo come una variabile numerica. Una volta eseguito lo *stemming*, si considerano tutti gli *stem* trovati nell'insieme di tutti i testi in esame, e per ogni singolo testo si costruirà un vettore detto *word stem profile*. Se nel testo è presenta lo *stem* nella corrispondente posizione del vettore si inserire un 1, altrimenti in caso di assenza si inserirà uno 0. Quindi per ogni testo $\mathbf{i}$ del corpus, viene associato uno *word stem profile* $\mathbf{S_i}$. Ogni testo apparterrà ad una categoria semantica $\mathbf{D_k}$, con $\mathbf{k} = \mathbf{1}, \ldots, \mathbf{K}$, dove $\mathbf{K}$ è il numero totale di categorie semantiche. Nella tabella 2.2 nella pagina successiva si riporta un esempio di una matrice di *stemming*.

- **Testo 1:** Questo libro è molto interessante;
- **Testo 2:** Dove hai comprato quel libro?;
- **Testo 3:** Il libro della Bibbia è il più comprato.
- **Testo 4:** Dovresti leggerlo questo libro!.

Testo	***libr***	***molt***	***interess***	***comprat***	***bibbia***	***legger***
T1	1	1	1	0	0	0
T2	1	0	0	1	0	0
T3	1	0	0	1	1	0
T4	1	0	0	0	0	1

Tabella 2.2: Esempio di una matrice di stemming

2.3 Il metodo di Hopkins e King

Sia $\mathbf{D} = (D_1, D_2, \ldots, D_K)$ l'insieme delle diverse categorie semantiche in esame. Sia $\mathbf{S_i} = (s_1, s_2, \ldots, s_M)$ l'insieme delle parole o *stem* $s_1, s_2, \ldots, s_M$, utilizzate in un testo per esprimere un *sentiment* tra quelle del vettore $\mathbf{D}$. Immaginiamo di dividere l'insieme dei testi in due gruppi: il *training set*, ovvero il sottoinsieme di testi che verrà codificato manualmente, e il sottoinsieme dei testi rimanenti di cui si vuole predire il contenuto semantico in modo automatico. Un gruppo di codificatori umani (*hand coders*) legge i test del *training set* e codifica manualmente questi testi assegnando una delle categorie $\mathbf{D_j}$, $j = 1, \ldots, K$. In questa fase di *training* del classificatore statistico, ovvero il *tagging* dei testi, vengono associate le parole $\mathbf{S_i}$ ai contenuti $\mathbf{D}$. A partire dalla matrice completa degli *stem* $\mathbf{S}$, si calcola la frequenza delle volte in cui una certa sequenza di *stem* $(s_1, s_2, \ldots)$ è associata ad una particolare categoria D_j. Ad esempio, quante volte la sequenza $\mathbf{s} = (s_1, s_2, s_3, \ldots, s_M) = (1, 1, 0, \ldots, 1)$ è associata ad una particolare categoria D_1?. Questa frequenza, sul totale dei testi analizzati N, è una stima delle probabilità con cui, prendendo un testo a caso dal *corpus*, osservata esattamente quella sequenza di *stem*, quel testo parla effettivamente della categoria scelta D_j. Questa è appunto la fase di *training* dell'algoritmo e produce una stima della probabilità $\mathbf{P}(D_j|\mathbf{s})$. Se si ripete la stessa operazione per tutte le sequenze $\mathbf{s}$ associate alla categoria D_j, possiamo stimare la probabilità complessiva $\mathbf{P}(D_j|\mathbf{S})$. Ciò vuol dire che, per ogni sequenza $\mathbf{s}$, avremo $\mathbf{P}(D_1|\mathbf{S})$, $\mathbf{P}(D_2|\mathbf{S})$, $\ldots$, $\mathbf{P}(D_K|\mathbf{S})$. Quindi, preso un testo a caso dal *test set* con vettore di *stem* $\mathbf{s} = \mathbf{S_i}$, il modello ci dirà le probabilità di associazione ad ogni categoria. Immaginando di aver classificato tutti gli N test del *corpus* in questo modo, cioè scegliendo di classificare un testo sulla base della categoria che riceve la probabilità più alta, possiamo passare dalla classificazione individuale di un testo alla classificazione aggregata semplicemente contando quante volte i testi sono classificati in tal modo. In termini matematici questo si scrive in modo matriciale come:

$$P(D) = P(D|S) \times P(S) \tag{2.1}$$

dove $\mathbf{P(D)}$ è la distribuzione aggregata delle opinioni, cioè un vettore di probabilità: $\mathbf{P}(D) = (\mathbf{P}(D_1), \mathbf{P}(D_2), \ldots, \mathbf{P}(D_K))$, di dimensione $K \times 1$. $\mathbf{P(S)}$ è un vettore di probabilità di dimensione $2^M \times 1$, in quanto ogni vettore di *stem* $\mathbf{s} = (s_1, s_2, s_3, \ldots, s_M)$ è una combinazione di M variabili

binarie. Infine $\mathbf{P(D|S)}$ è una matrice di dimensioni $K \times 2^M$.

Mentre non vi è alcun problema nella previsione individuale della categoria D_j, una volta che le previsioni vengono aggregate per ottenere $\mathbf{P(D)}$ l'errore di ogni singola predizione prodotto dal classificatore statistico può amplificarsi.
Hopkins & King (2010) propongono un metodo che evita di passare dalla classificazione individuale. Prendiamo come $\mathbf{P(S)}$ la distribuzione degli *stem* dell'intero insieme di dati (*training set* e *test set*), possiamo scomporla come:

$$P(S) = P(S|D) \times P(D) \tag{2.2}$$

La matrice $\mathbf{P(S|D)}$ rappresenta la probabilità che una particolare sequenza di *stem* $\mathbf{s}$ di $\mathbf{S}$ compaia all'interno dei testi che sono classificati secondo una particolare categoria D_j. Questa matrice $\mathbf{P(S|D)}$ ha dimensione $2^M \times K$ e la relazione va intesa come:

$$P(S = s) = \sum_{j=1}^{K} P(S = s|D = D_j) \times P(D_j) \tag{2.3}$$

al variare di $\mathbf{s}$ in $\mathbf{S}$. Poichè è nota la distribuzione $\mathbf{P(S|D)}$ solo per i testi del *training set*, cioè quelli effittivamente codificati, si deve fare l'ipotesi che il linguaggio utilizzato nel *training set* (**TR**) nella discussione dell'opinione D_j sia lo stesso adoperato da tutti i testi del *corpus*, ovvero che:

$$P_{TR}(S|D) = P(S|D) \tag{2.4}$$

Sarà quindi possibile riscrivere l'equazione (2.2) come:

$$P(S) = P_{TR}(S|D) \times P(D) \tag{2.5}$$

A questo punto la soluzione del problema si ottiene invertendo la 2.5, ovvero:

$$P(D) = P_{TR}(S|D)^{-1} \times P(S) \tag{2.6}$$

Il vantaggio fondamentale di questo approccio di inversione è che non viene utilizzata la classificazione individuale e successivamente aggregata ma viene stimata direttamente la distribuzione aggregata $\mathbf{P(D)}$ riducendo il margine di errore.

2.3.1 Selezione delle categorie

L'individuazione delle categorie da analizzare, può risultare un problema banale e di facile risoluzione. Occorre però sottolineare che le categorie scelte devono soddisfare tre requisiti fondamentali: esclusività, esaustività e omogeneità. Quindi si introdurranno, oltre alle classiche categorie di sentiment positivo e negativo, anche la possibilità di un sentiment neutro (espressione di un sentiment neutrale) e di un sentiment *no opinion* (nessuna opinione anche se trattano il tema in esame). In questo caso di studio, come in molte altre applicazioni dell'analisi testuale, è fondamentale l'introduzione di una categoria (*off-topics*) che possa contenere tutti quei testi che non trattano l'argomento in esame. Quest'ultima categoria, come vedremo, risulterà essere quella più frequente, infatti eseguendo il *download* dei testi per parole chiave, si sono presi molti testi che contengono una o più di queste parole ma che non trattano in realtà il tema di studio. La tabella 2.3 riporta le categorie scelte.

Categoria	Nomi delle categorie	Descrizione
1	Positivi	Commenti espressamente positivi
2	Negativi	Commenti espressamente negativi
3	Neutri	Commenti che non si sbilanciano né sul negativo né sul positivo
4	No Opinion	Commenti inerenti ma che non esprimono un giudizio
5	Off Topics	Commenti fuori tema

Tabella 2.3: Categorie solite per la *Sentiment Analysis*

2.3.2 Hand-Coding

Come specificato nella trattazione del metodo di Hopkins e King, occorre anteporre al metodo automatico di calcolo delle distribuzione nelle varie categorie, una fase di *training* dell'algoritmo. Si sono dunque costruiti due *training set* (uno relativo a Facebook e l'altro relativo a Twitter), dove ogni singolo *post* o *tweet* è stato classificato manualmente da un *coder*. Questa procedura è fondamentale e richiede la miglior accuratezza possibile, infatti come si è visto prima, si supporrà nel seguito che $P_{TR}(S|D) = P(S|D)$, e dunque i risultati ottenuti saranno dipendenti dalle scelte eseguite in questa fase. Per questi motivi, è opportuno che chi esegue la fase di *hand-coding* deve avere una buona conoscenza del problema in esame, delle categorie possibili e minimizzare il dubbio durante l'associazione testo - categoria.

Pur applicando queste precauzioni, si rischia di cadere in diversi problemi di difficile risoluzione. La soggettività nell'assegnazione permette la possibilità di una diversa classificazione a seconda del *coder*. Infatti per un individuo, il documento **i** può essere assegnato alla categoria **j** mentre per un'altra persona potrebbe appartenere alla categoria $\mathbf{j}'$. Per rendere meno influente la soggettività nell'assegnazione delle categorie in fase di *tagging* manuale, si è scelto di far etichettare i testi a tre *coders* differenti. Nel testi in cui i diversi *coders* non sono d'accordo, verranno successivamente discussi per arrivare ad una categoria di assegnazione comune. Nel caso questo sia impossibile, è preferibile rimuovere quel testo dal *training set*. Verranno successivamente introdotte delle metriche sul grado di accordo tra i diversi *coders* che hanno eseguito la procedura di *training*. Queste ci permetteranno di valutare a priori l'affidabilità delle analisi svolte. Ovviamente, in termini di tempo, questa procedura è la fase più gravosa dell'intero metodo, ma occorre rimarcare come, se è presente indecisione già nelle battute iniziali da *coders* umani, non ci si possono aspettare risultati migliori ottenibili da un classificatore automatico.

Contrariamente ad altri metodi statistici, in questo caso non è possibile stabilire un regola che permetta di scegliere a priori la dimensione del *training set*, ovvero il numero minimo di testi da etichettare affinché si posso ottenere una stima sufficientemente precisa del modello. Ovviamente la dimensione del *training set* dipenderà dal numero di categorie scelte e dalla lunghezza dei testi in esame.

2.3.3 Ipotesi e loro verifica

Hopkins e King prevedono come unica ipotesi alla base del loro metodo, la rappresentatività linguistica dei testi del *training set* rispetto a tutto il *corpus* di testi, ovvero in termini matematici, che sia verificata l'assunzione descritta dall'equazione 2.7

$$\mathbf{P_{TR}(S|D) = P(S|D)} \tag{2.7}$$

Questa condizione è verificata quando nel *training set* sono presenti un numero sufficiente di *stem profile*, tali da poter individuare e caratterizzare ogni categoria. Generalmente questa ipotesi può venire meno quando:

- La fase di *training* dell'algoritmo è stata eseguita troppo tempo prima rispetto al *download* dei dati in esame (risulta difficile pensare di applicare il metodo ai moderni SMS utilizzando come *training set* la corrispondenza postale);

- Il *training set* non è stato generato in modo casuale ma si è catturato solo uno specifico modo di esprimere i giudizi (per esempio se si sceglie come *training set* i testi prodotti dai solo ragazzi rispetto ad una classe più eterogenea).

Nel caso di un'eventuale evoluzione e quindi una conseguente trasformazione del linguaggio, occorrerà aggiornare il *training set* con nuovi testi che ben identificano questi cambiamenti.

Non sarà quindi richiesto che ciascuna categoria sia presente con lo stesso numero di testi nel *training set*, ma che sia linguisticamente rappresentata. Qualora le categorie scelte a priori risultassero non rappresentative all'interno del *training set*, si dovranno ridefinire la selezione delle categorie.

2.3.4 Bagging

Nell'equazione 2.6 a pagina 14 si utilizza una matrice molto grande, dunque il costo computazionale nell'eseguire il metodo è elevato. Inoltre il numero di testi presenti nel *training set* è molto minore del numero totale di documenti, perciò il numero totale di *stem profiles* utilizzati per stimare $\mathbf{P}(\mathbf{S}|\mathbf{D})$, è decisamente minore rispetto al numero totale dei possibili *stem profiles* ottenibili da tutti gli *stem* presenti nel testo.
Quest'ultimo problema comporta delle probabilità molto piccole e di conseguenza ad una matrice sparsa. Per ridurre i costi computazionali e la sparsità si utilizzata il metodo Monte Carlo proposto da King & Lu (2008) denominato come ***bagging***.

L'operazione consiste nell'effettuare il metodo di Hopking e King utilizzando dei sottoinsiemi di *stem* ridotti costituiti quindi da un numero minore di stem (solitamente variabile tra i 5 e i 25 *stem*). Successivamente si ottengono per ognuno dei sottoinsiemi le stime di $\mathbf{P}(\mathbf{D})$ che vengono mediate aritmicamente.

Il metodo di selezione degli *stem* da considerare in ogni sottoinsieme diventa a questo punto fondamentale. Nel loro metodo, Hopkins e King, selezionano gli *stem* in modo casuale, assegnando ad ognuno la stessa probabilità di essere pescato. Ciò non toglie che in alcune casistiche sia meglio assegnare agli *stem* dei pesi differenti a seconda per esempio delle frequenze di presenza all'interno del *corpus* dei testi.

2.3.5 Vantaggi e svantaggi

Riassumendo, il metodo di Hopkins e King, risulta essere facilmente applicabile. Ciò non comporta una bassa profondità dei risultati, anzi le capacità del metodo risultano sorprendenti. In questo caso, a differenza degli altri metodi di analisi testuale, non si richiedere la conoscenza o la selezione di particolari *stem* da individuare nel *corpus* ma il metodo introdotto è capace in maniera automatica di identificare gli *stem "portatori"* di *sentiment*, anche quelli che probabilmente sarebbero potuti sfuggire.

Il vantaggio fondamentale del metodo va ricercato nell'utilizzo dei vari *stem*. Poiché non vengono più utilizzati come entità isolate, ma come componenti di un insieme di parole (lo *stem profile*), si riescono a catturare combinazione di più *stem*, avvicinandoci maggiormente al modo in cui anche un essere umano *"percepisce"* il *sentiment* di un testo.

Lo svantaggio principale, come già discusso precedentemente, consiste nella fase di *training* dell'algoritmo. La procedura di *hand-coding* risulta essere sia molto dispendiosa in termini di tempo, sia molto delicata per la bontà dei risultati. Infatti, ricordando l'ipotesi espressa nella 2.7 a pagina 17 alla base del metodo, si evince l'importanza dell'operazione di *tagging manuale*.

Infine, occorre ripetere, che il metodo, se pur integrato con la fase di *bagging*, presenta dei costi computazionali molto elevati (basti pensare al numero di testi analizzati).

Capitolo 3

Case Study: Gli atenei italiani

Negli ultimi anni il modo di pensare al *business* nel mondo delle aziende è cambiato. Tutte le più grandi aziende del mondo sono alla costante ricerca di nuovi modi per migliorare le loro decisioni chiave, anche facendo leva su quello che i proprio clienti condividono. Con questo nuovo approccio, i canali digitali e i social media in particolare, certamente rappresentano un'opportunità di raccogliere dati e aspettative dei clienti.
Anche il mondo dell'università, se pur in maniera diversa, non è esente da queste logiche di mercato e di *governance*. E' di fondamentale interesse capire quali giudizi gli studenti esprimono nei confronti del proprio ateneo di appartenenza, in modo tale da individuare i punti critici nella gestione dell'università.

Figura 3.1: *Sentiment and Opinion Analysis*: dai testi ai *report*

Si vuole quindi sfruttare il metodo descritto nei capitoli precedent per individuare e studiare tale *sentiment*. Lo studio si è sviluppato sui due fronti principali:

- **Facebook:** Post scaricati dal 1/1/2014 al 30/06/2015
- **Twitter:** Post scaricati dal 1/1/2015 al 30/06/2015

La 3.2 riporta le numerosità in esame.

Post totali:	**97595**	**Tweet totali:**	**56394**
Post istituzionali:	**5508**	**Tweet istituzionali:**	**9331**
Post non istituzionali:	**92087**	**Tweet non istituzionali:**	**47063**

Figura 3.2: I dati dai due social network

I *software* utilizzati in questo studio, sono tutti *freeware* (licenza e *download* gratuiti). In particolare l'applicazione del metodo e la visualizzazione grafica dei risultati sono stati eseguiti sul *software* di analisi statistica R (2013) sui sistemi operativi *Windows 7* e *Ubuntu*.

3.1 Analisi

3.1.1 Categorie

Come riportato nelle sezioni precedenti, sono state individuate delle categorie di *sentiment* per ogni ogni tweet e post analizzato. Queste categorie sono state scelte anche per lo studio dell'*Opinion Analysis* (la fase in cui si analizza la natura del giudizio espresso). Occorre ora definire le categorie, sia per quanto riguarda la *Sentiment Analysis* sia per l'*Opinion Analysis*, più specificatamente al caso di studio. La tabella 3.1 riporta le classi scelte per la *Sentiment Analysis* dei post/tweet relativi alle università. L'*Opinion Analysis* invece mira ad un'analisi più profonda indagando i motivi che hanno spinto gli studenti ad esprimere un determinato giudizio. Nella tabella 3.2 nella pagina seguente si riportano le categorie scelte. Come si può vedere vengono esplose le categorie positive e negative della *Sentiment Analysis*, mentre le categorie in cui non è presente giudizio o fuori tema vengono accorpate nella categoria *Altro*.

Sentiment Analysis	
Nomi delle categorie	**Descrizione**
Positivi	Commenti espressamente positivi sul singolo ateneo o sull'intero sistema universitario
Negativi	Commenti espressamente negativi sul singolo ateneo o sull'intero sistema universitario
Neutri	Commenti riguardo l'università che non si sbilanciano né sul negativo né sul positivo
No Opinion	Commenti che non esprimono un giudizio ma che riguardano comunque le università
Off Topics	Commenti che non riguardano l'università

Tabella 3.1: Categorie *Sentiment Analysis* del *Case Study*

Opinion Analysis	
Sentiment	**Nomi e descrizione delle categorie**
Positivi	**Generico**: Sentiment positivi nei confronti dell'università o del sistema universitario espresso in modo esplicito ma non riconducibili alle altre categorie positive, tra cui ad esempio sport e trasporti pubblici per raggiungere gli atenei.
	Didattica & Student Experience: Giudizi positivi relativamente a esami, corsi, programmi, docenti, lezioni, esercitazioni, libri.
	Supporto alla didattica: Commenti positivi sulla gestione e il funzionamento del sito, della segreteria, di eventi/attività organizzati dall'ateneo (convegni, visite, conferenze, ...) e delle collaborazioni con altre università.
	Sicurezza persone e asset fisici: Adeguatezza delle infrastrutture e dei controlli.
	Amministrazione e Governance: Amministrazione e Governance giudizi positivi su decisioni di policy dell'ateneo tipo cambiamento di lingua inglese o altre modifiche riguardanti rettori, personale, tasse.
Negativi	**Generico**: Critiche esplicite al sistema universitario, al singolo ateneo, alla vita da studente o altri commenti che non appartengono alle altre categorie negative, tra cui sport e trasporti pubblici per raggiungere gli atenei.
	Didattica & Student Experience: Giudizi negativi relativamente ad esami, corsi, programmi, docenti, lezioni, esercitazioni, libri e tesi.
	Supporto alla didattica: Commenti relativi a malfunzionamenti e cattiva gestione del sito, delle segreterie e degli eventi/attività organizzati dall'ateneo e delle collaborazioni con altre università.
	Sicurezza persone e asset fisici: Non adeguatezza delle infrastrutture e dei controlli.
	Amministrazione e Governance: critiche negative su decisioni di policy dell'ateneo tipo cambiamento di lingua inglese o altre modifiche riguardanti rettori, personale, tasse.
Altro	**Neutri, No Opinion e Off Topics**: Tutto ciò che non esprime un giudizio sugli atenei come eventi non relativi all'università o testi non comprensibili.

Tabella 3.2: Categorie *Opinion Analysis* del *Case Study*

3.1.2 Tagging manuale

Per minimizzare la soggettività del singolo valutatore, nella fase di *hand-coding*, si sono adoperati tre diversi *coders*. Ogni singolo *coder*, in separata sede dai rimanenti due, ha proceduto ad etichettare ogni singolo post o tweet del training set. Nella tabella 3.3 si riporta il grado di accordo tra i tre diversi *coders* presi singolarmente e considerandoli tutti insieme sia per il *training set* di Facebook che per quello di Twitter.
Come misura di accordo tra le coppie di *coders* (*inter-rater reliability*) è stata utilizzata la *Kappa* di Cohen (1960), ovvero un coefficiente statistico atto a rappresentare il grado di accuratezza di una classificazione statistica. Tale indice confronta gli accordi effettivamente osservati con quelli che invece si sarebbero verificati solo per caso.
Mentre nel calcolo del grado di accordo tra più di due *coders* si è adoperato l'indice statistico denominato *Kappa* di Fleiss (1971). Quest'ultimo può essere interpretato come espressione della misura in cui il grado di accordo osservato tra i valutatori superi quello che ci si aspetterebbe se tutti i *coders* avessero compiuto il loro *tagging* in modo casuale.

Come si può notare, la concordanza tra le coppie di *coders* è decisamente maggiore nel caso di Facebook. Probabilmente, questo fenomeno, è dovuto alla lunghezza dei testi nei due *social network*. Infatti nel caso di Facebook il testo è mediamente più lungo e quindi può contenere più informazione che aiutano il *coder* ad etichettarlo.

	Concordanza Coders			
	C1 vs C2	**C1 vs C3**	**C2 vs C3**	**Fleiss**
Facebook	74,72%	81,75%	88,89%	81,72%
Twitter	37,85%	43,12%	33,25%	38,93%

Tabella 3.3: Grado di accordo tra i diversi *Coders*

Successivamente, nei casi dubbi, ovvero in quei casi dove tutti e tre i *coders* etichettavano il testo in maniera differente, si è proceduto ad una nuova discussione della procedura di *tagging*, cercando un accordo su quale categoria scegliere. Infatti, ricalcolando il grado d'accordo tra tutti e tre i

coders, si ottiene **96,23%** per Facebook e **69,00%** per Twitter, migliorando così l'affidabilità della fase di *hand-coding*.

3.1.3 Misura di accuratezza

Introduciamo in questa sezione l'indice di accuratezza che utilizzeremo per valutare la bontà del metodo utilizzato. Per procedere a tali verifiche si attuerà un processo di *cross-validation*, dividendo l'insieme dei testi etichettati manualmente in *training set* e *testing set*. E' stata scelta come percentuale di divisione il 75%, ovvero il 25% dei testi manualmente etichettati sarà utilizzato come sottoinsieme su cui testare il classificatore, e il 75% per identificare e costruire il classificatore.

Sia ora $\mathbf{T_j}$ la percentuale vera di testi nel *testing set* nella categoria $\mathbf{j}$, e $\mathbf{E_j}$ la percentuale stimata dal classificatore. Si introduce allora l'indice comunemente adoperato in letteratura:

$$I_{HK} = \sqrt{\frac{\sum_{j=1}^{J} (T_j - E_j)^2}{J}} \tag{3.1}$$

L'indice appena introdotto è una distanza tra due vettori di probabilità. La stima sarà tanto più accurata quanto più il valore di $\mathbf{I_{HK}}$ sarà prossimo allo 0.

3.1.4 Il pacchetto *ReadMe*

Il metodo presentato nelle sezioni precedenti è in alcune sue fasi sviluppato e disponibile nel pacchetto *ReadMe* Hopkins et al. (2013) su *R*. Un pacchetto è uno *"scatolone"* contenente diversi funzioni. In questo caso, il *package ReadMe*, interfacciabile con *Python*, presenta funzioni che, prendendo come *input* i testi, le etichette e vari parametri, restituisce come *output* la distribuzione delle probabilità in ciascuna delle categorie. I parametri di ingresso sono:

- **Threshold:** All'interno dell'intero *corpus*, gli *stem* che compaiono troppo frequentemente (come ad esempio *ciao*) o troppo raramente non saranno discriminanti nella scelta del *sentiment* da associare a quel particolare tweet o post. Si impone così una soglia (*threshold*) sotto o sopra la quale gli *stem* troppo o poco frequenti vengono eliminati. Nel nostro caso si è deciso di lasciare il valore di *default* impostato nel pacchetto *ReadMe* dell'**1%**.

- **Subset:** Nella fase di *bagging*, ovvero la procedura per snellire l'elevato costo computazionale, i vari *stem* (che hanno superato il limite imposto dalla *threshold*) vengono ripartiti in n sottoinsiemi diversi e conseguentemente viene calcolata la stima di P(D) (che andrà poi mediata aritmeticamente su tutti i sottoinsiemi). Ovviamente maggiore è il numero imposto di sottoinsiemi e maggiore sarà l'accuratezza della stima. Nel nostro caso si è deciso di lasciare il valore di *default* impostato nel pacchetto *ReadMe* di **300** sottoinsiemi.

- **Features:** Il numero di *stem* (*features*) da utilizzare per la costruzione di ogni singolo *subset*. Per la scelta di questo parametro si può procedere, seguendo il suggerimento di Hopkins e King, per cross-validazione. Si confronterà la scelta di 15 *features* (scelta di *default* del pacchetto) per sottoinsieme con quello di 8 *features* per gruppo, minimizzando l'indice I_{HK} presentato precedentemente. Eseguendo l'analisi sull'intero dataset di dati si sono ottenuti i valori $I_{HK}^{8} = 0.32$ e $I_{HK}^{15} = 0.13$. Si è dunque scelto di impostare il numero di *features* ad **15**.

3.2 Twitter

La prima fase dell'analisi si è sviluppata considerando i *tweet*, e quindi analizzando il social network Twitter. I *tweet* analizzati coprono un arco temporale che va dal primo Gennaio 2015 al 30 Giugno 2015. Nella tabella 3.4 vi è una prima analisi descrittiva del dataset di Twitter.

Numero tweet totali	56394
Numero tweet istituzionali	9331
Numero tweet non istituzionali	47063
Numero di atenei analizzati	22
Numero medio di tweet per ateneo	∼2560

Tabella 3.4: Analisi descrittiva dataset Twitter

La divisione dei *tweet* nei vari atenei è raffigurata in figura 3.3. Si può notare come le numerosità siano molto diverse per singola università e come in linea generale il numero di *tweet* istituzionali sia molto minore di quelli non istituzionali.

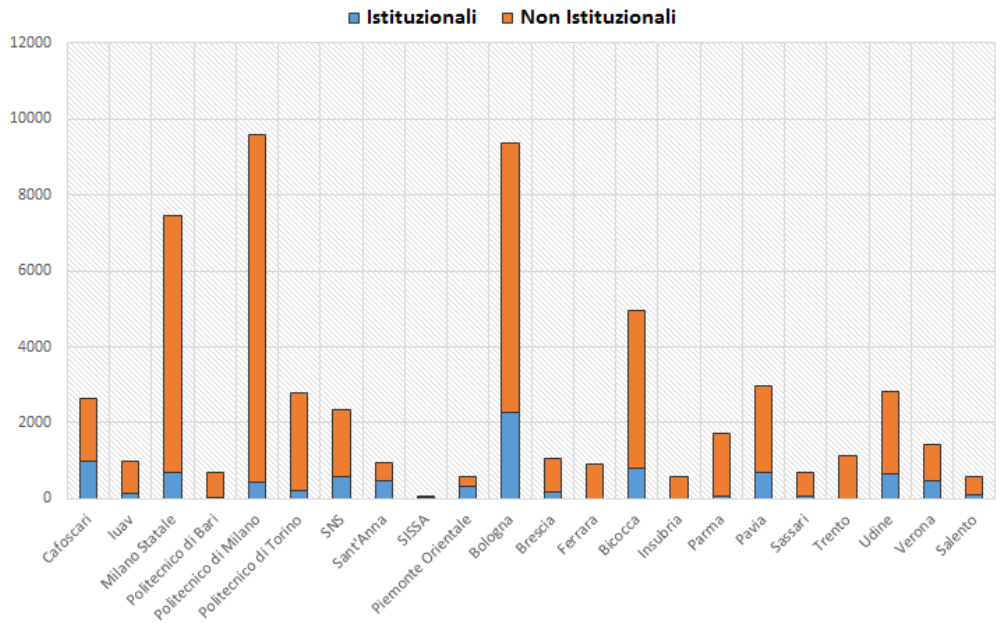


Figura 3.3: Distribuzione dei *tweet* per ateneo e per istituzionalità

3.2.1 Il training set e l'hand-coding

L'analisi è iniziata con la costruzione del *training set* e la seguente fase di *hand coding*. Il *training set*, formato da 1206 *tweet*, è stato costruito cercando di bilanciare il numero di tweet nei diversi mesi e nei diversi atenei, ovvero di ottenere per ogni ateneo e per ogni mese dell'anno lo stesso numero di *tweet* nel *training set*.

Come si evince dalla figura 3.4, che mostra la distribuzione dei *tweet* nel *training set*, in fase di *hand-coding* non si sono riscontrati *tweet* appartenenti alla categoria **Neutra**. Ovviamente, non avendo nessun riscontro in fase di addestramento in questa categoria, il classificatore automatico la lasciera vuoto. Si è deciso dunque di eliminare questa categoria.

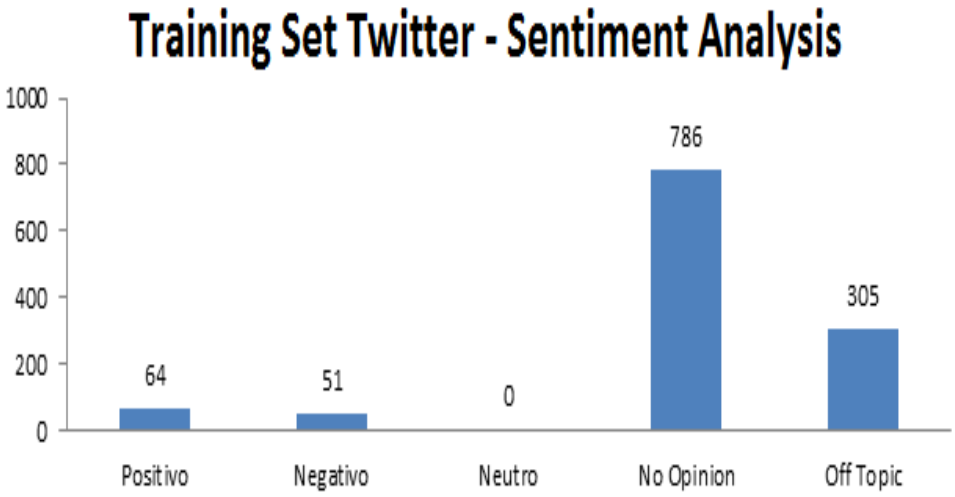


Figura 3.4: Numero *tweet* per le categorie di *sentiment analysis* nel *training set*

Durante la fase di *tagging* manuale per la *sentiment analysis*, si è proceduto in simultanea, al *tagging* per l'*opinion analysis*. I risultati ottenuti dall'*hand-coding* sono riportati in figura 3.5 nella pagina successiva.

Come si può notare, la maggior parte dei *tweet* presenti nel *training set*, sono raccolti dalla categoria **Altro**, ovvero la categoria dell'*opinion analysis* che include le categorie di **No Opinion** e **Off Topics** della *sentiment analysis*.

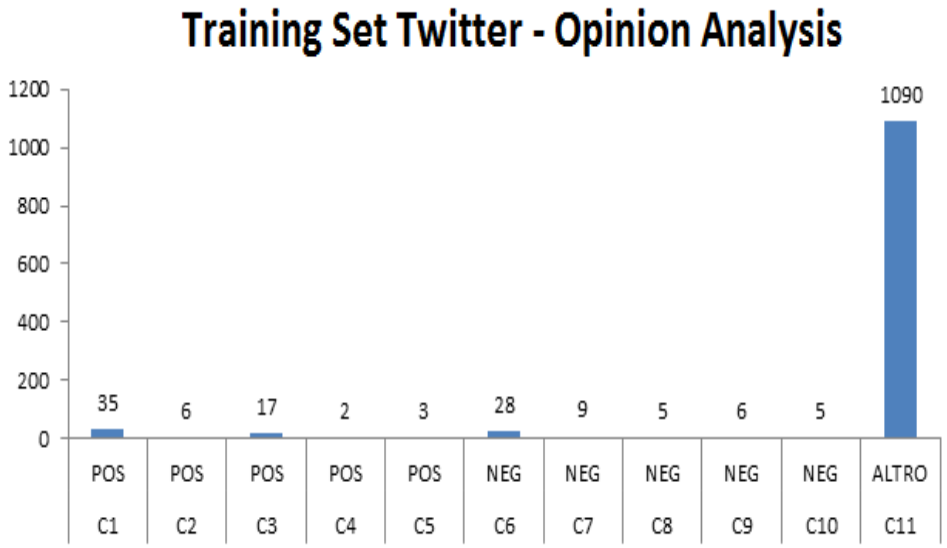


Figura 3.5: Numero *tweet* per le categorie di *opinion analysis* nel *training set*

Questo risultato è in accordo con le informazioni a priori sulla modo di comunicare sui *social network* e sulle modalità di *download* dei *tweet*. Nonostante la bassa numerosità di alcune categorie di *opinion analysis*, si è deciso comunque di conservale distinte e non accorparle, per non perder l'informazione relativa ad ogni determinata categoria.

3.2.2 Elaborazione dei *tweet*

Proseguendo lungo le linee guida del metodo di Hopkins e King, dopo la fase iniziale di *tagging* manuale, a questo punto si procede con il *preprocessing* dei testi, ovvero la fase di pulizia e *stemming* dei singoli *tweet*. Viene inizialmente uniformato il testo in minuscolo e rimossi gli spazi bianchi in eccesso tra le diverse parole. Poi è stato necessario eliminare i collegamenti *URL* a siti internet presenti in moltissimi *tweet*. Successivamente, utilizzando le *stopword* riportate in Appendice B, si sono eliminate quelle parole che non portano contenuto semantico nel testo e infine eliminata la punteggiatura.

Occorre precisare, che la limitazione dei 140 caratteri imposti da Twitter, ha portato ad un linguaggio diverso (e di qui la necessità di due *training set* diversi per i due *social network*) ricco di abbreviazioni (come ad esempio

cmq o *xkè* per dire *comunque* o *perchè*) e in cui si rimarca un determinato *sentiment* anche solo prolungando una parola (per esempio *bravaaaa*).
Mentre nel caso di parole abbreviate è possibile una correzione aggiungendole nella lista delle *stopwords*, per le parole *"prolungate"*, si ha una perdita di informazioni, infatti non sarà possibile riconoscore lo *stem* (lo *stem brav* sarà analizzato separatamente da quello *bravaaaa*).

Occorre ricordare che l'ordine di esecuzione dei singoli passi è fondamentale. Infatti l'inversione di due o più *step* può portare a risultati completamente diversi. L'ultima operazione sui testi è la fase di *stemming*, apportata utilizzando lo *stemmer* di Porter, presente nel *package tm* scaricabile in *R*.

Nelle prossime sezioni, si riporteranno dei grafici riassuntivi dei risultati ottenuti dalle diverse analisi (*sentiment* e *opinion*).

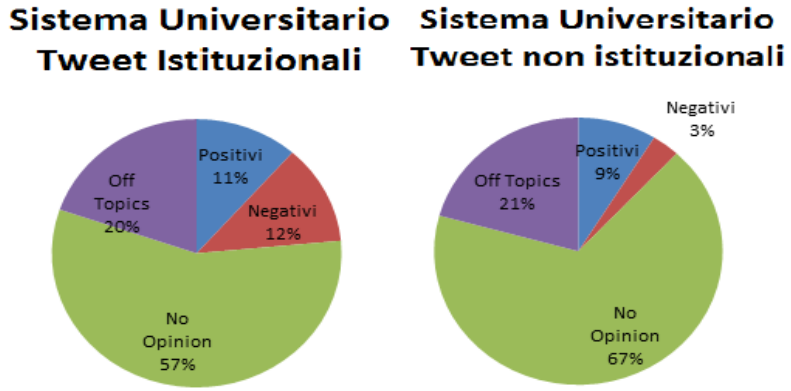
3.2.3 Sistema Universitario

Utilizzando le configurazioni introdotte precedentemente, si è proceduto all'esecuzione della *sentiment analysis* e dell'*opinion analysis* dei *tweet* per l'intero sistema universitario.

Sentiment Analysis

La figura 3.6 a fronte mostra, per l'intero sistema universitario, sia nel caso di *tweet* istituzionali e non, la distribuzione degli stessi nelle quattro categorie analizzate.

Focalizzando l'attenzione sui soli *tweet* che esprimono giudizio, ovvero solo quelli positivi o negativi, si può notare dalla figura 3.7 nella pagina successiva come le percentuali di tweet negativi e positivi, sia presi singolarmente che complessivamente, sono maggiori nei *tweet* istituzionali.
Probabilmente ciò è dovuto al fatto che se uno studente che vuole esprime un giudizio positivo, o sopratutto un disappunto, utilizzerà l'*hashtag* istituzionale dell'ateneo, in modo da assicurarsi venga letto dalle persone competenti. Ovviamente occorre sempre considerare anche le diverse numerosità dei due campioni.



(a) Tweet istituzionali **(b)** Tweet non istituzionali

Figura 3.6: Distribuzione dei *tweet* per le categorie di *sentiment analysis* per l'intero sistema universitario

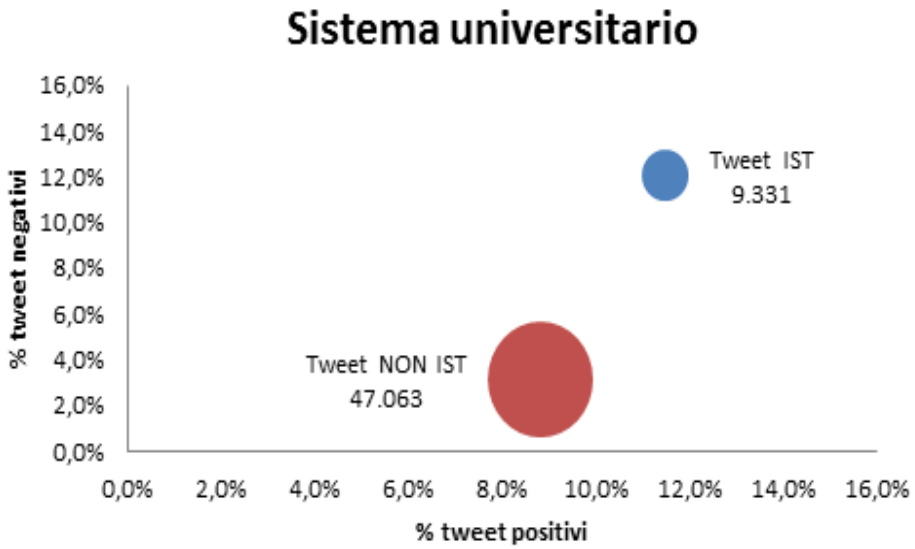


Figura 3.7: Confronto *tweet* con giudizio per l'intero sistema universitario per la *sentiment analysis*

Opinion Analysis

In questa sezione mettiamo in evidenza la composizione dei giudizi negativi e positivi rispetto alle cinque sottocategorie dell'opinion analysis:

- Generico;
- Didattica & student experience;
- Supporto alla didattica;

- Sicurezza persone e asset fisici;
- Amministrazione e Governance.

Come si vede dalla figura 3.8, nel caso di giudizi positivi, la maggior parte dei *tweet* ricadono nelle categorie **Generico** e **Supporto alla didattica**.
Gli aspetti più criticati, ovvero le categorie che presentano un maggior numero di *tweet*, sono il **Generico** e la **Sicurezza persone e asset fisici**.

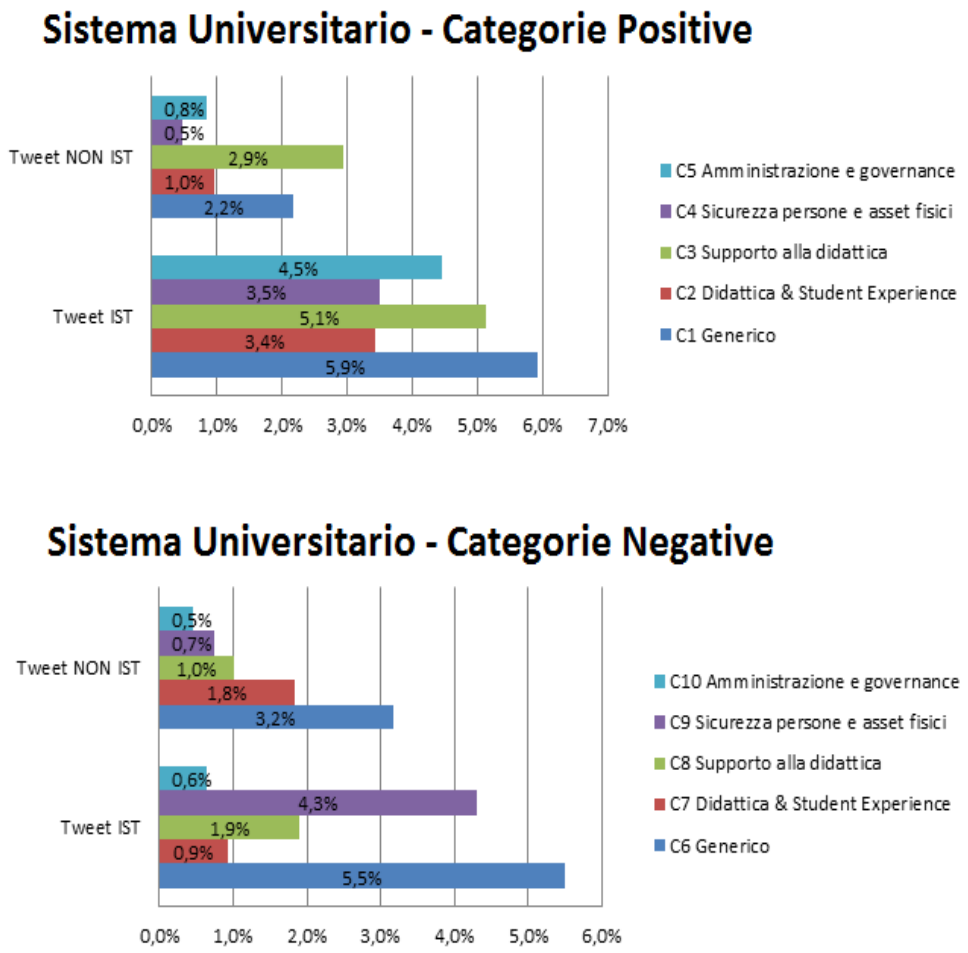


Figura 3.8: Confronto *tweet* con giudizio per l'intero sistema universitario per l'*opinion analysis*

3.2.4 Confronto tra gli atenei

Vogliamo adesso analizzare le differenze ricavate dalla *sentiment analylis* e dall'*opinion analysis* tra ogni singolo ateneo. Occorre sottolineare che per quanto riguarda i *tweet* istituzionali mancano le università di Ferrara, Insubria e Trento per nel caso non istituzionale si nota l'assenza della Sissa.

Ranking

Si vuole ora stilare una classifica che permetta di individuare gli atenei in cui gli studenti esprimono maggior disappunto. Si è costruito un indice sintetico dato dal rapporto tra la percentuale dei *tweet* negativi e la percentuale dei *tweet* con giudizio (quindi negativi e positivi). Ovvero l'indici di ranking sarà:

$$r_{neg} = \frac{\%TweetNegativi}{\%TweetNegativi + \%TweetPositivi}$$

Come si può notare dai risultati mostrati in figura 3.9, gli atenei occupano posizioni diverse a seconda se si considerano tweet istituzionali o meno. Nel primo caso si ha che l'ateneo con un *sentiment* negativo maggiore è l'**Università degli Studi del Piemonte Orientale**, mentre nel caso dei *tweet* non istituzionali si ha che l'**Università degli studi di Pavia** presenta un r_{neg} più elevato.

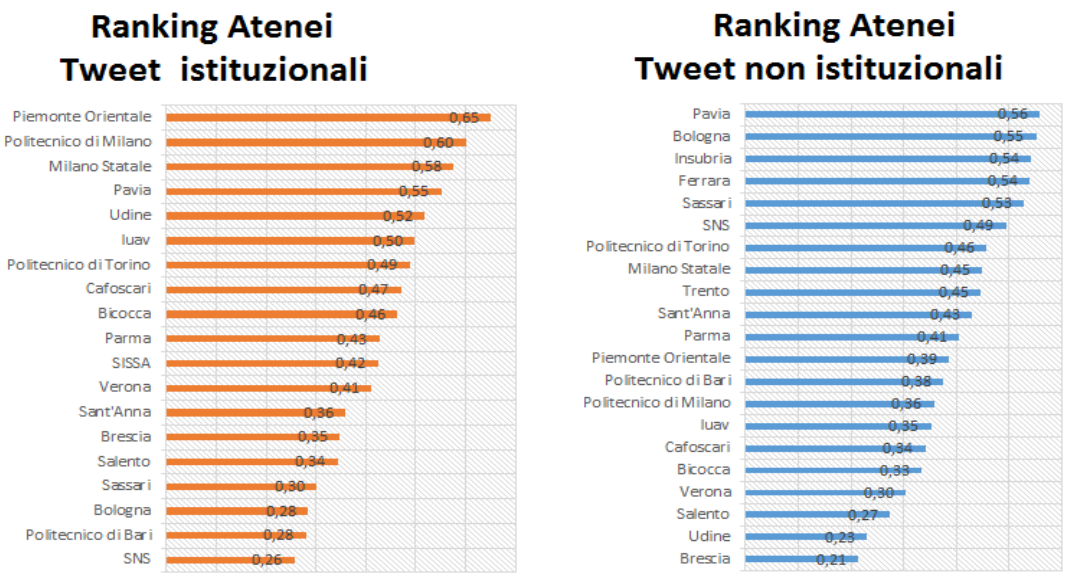


Figura 3.9: Ranking *tweet* con giudizio tra i diversi atenei per la *sentiment analysis*

Benchmarking

Si vuole adesso confrontare tra loro i diversi atenei in un grafico a bolle. Ciò permetterà di visualizzare graficamente anche il numero di *tweet* per ogni singola università, indice anche di affidabilità del risultato ottenuto (più è grande il campione di partenza e migliore sarà la stima ottenuta). Per evitare sovraffollamento e quindi rendere il grafico illeggibile, si sono selezionati gli atenei con almeno 200 *tweet* nel caso istituzionale e con almeno 1000 *tweet* nel caso non istituzionale.

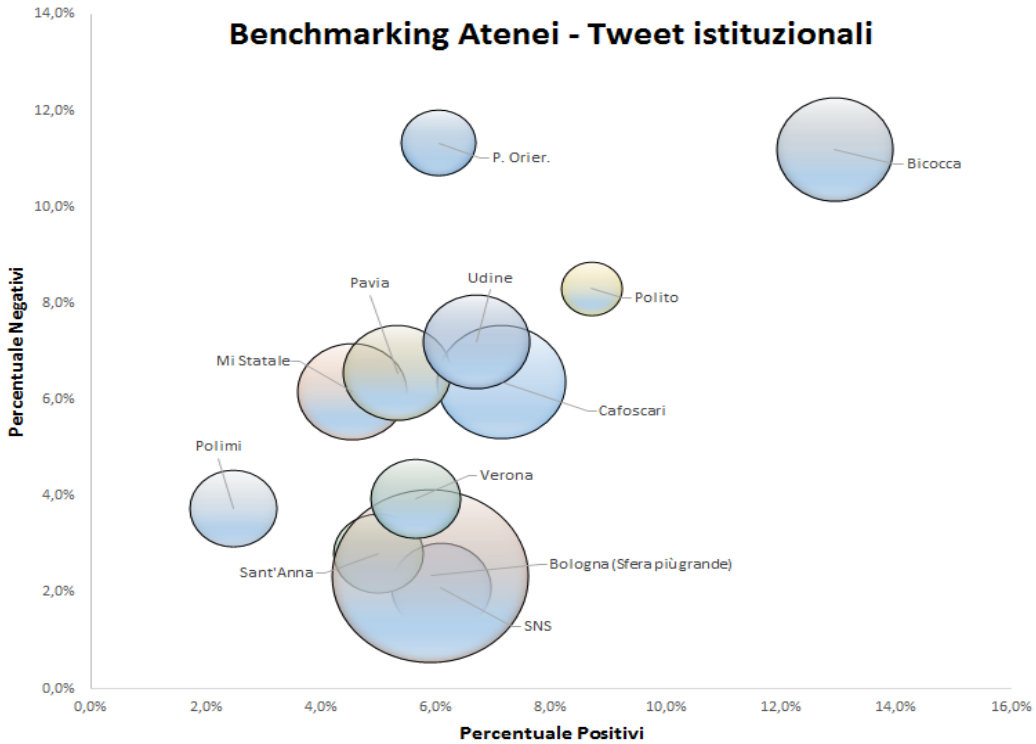


Figura 3.10: Benchmarking *tweet* istituzionali con giudizio tra i diversi atenei per la *sentiment analysis*

Dai grafici del benchmarking si possono ottenere dei rapidi confronti visivi tra le varie università. Nel caso istituzionale (figura 3.10), si nota la presenza di alcune università suddivise in due gruppi di università (il primo formato da *Milano Statale, Pavia, Udine e Cafoscari* e il secondo costituito da *Verona, Sant'Anna, Bologna e SNS*) che hanno caratteristiche simili.

Mentre nel caso non istituzionale (figura 3.11 nella pagina successiva) sembra esserci un unico gruppo con caratteristiche simili (costituito da *Polito, Trento, Milano Statale*).

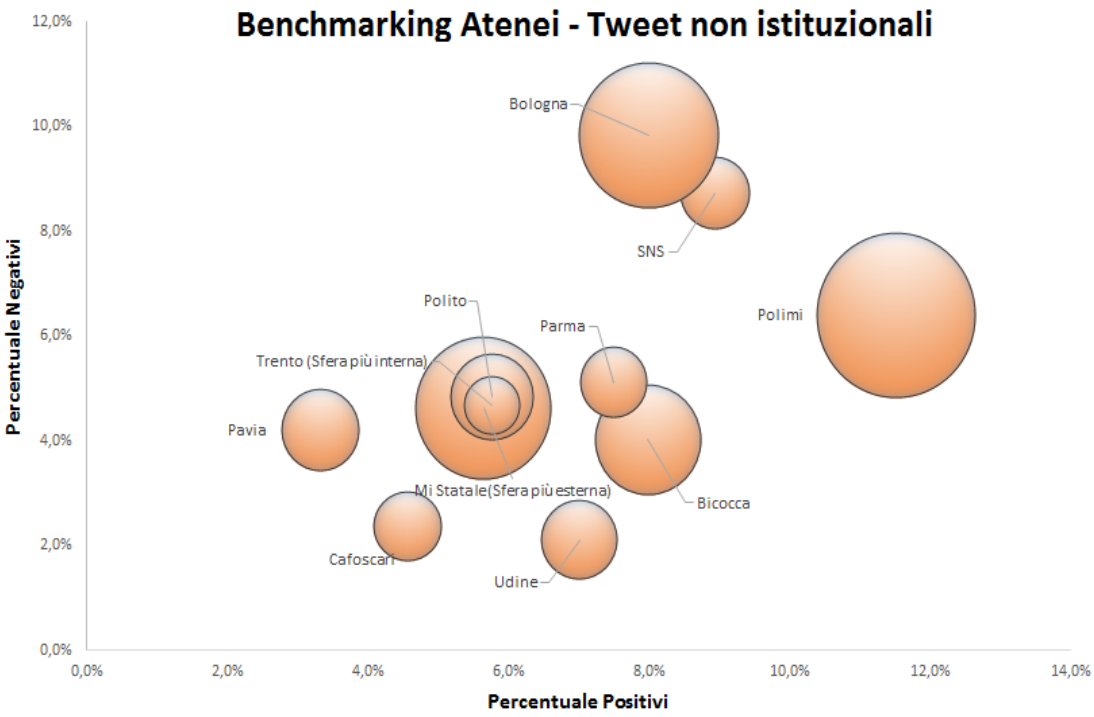


Figura 3.11: Benchmarking *tweet* non istituzionali con giudizio tra i diversi atenei per la *sentiment analysis*

Opinion Analysis

Per mostrare i risultati relativi all'*opinion analysis*, abbiamo scelto di utilizzare delle tabelle, in quanto permettono di confrontare tutte le categorie insieme, cosa che per leggibilità non sarebbe possibile producendo un grafico per ognuna delle cinque sottocategorie. In questo modo riduciamo tutto a solo due tabelle, dove le percentuali sono calcolate rispetto rispetto al totale dei tweet dell'ateneo.

Nella tabella 3.5 nella pagina seguente e nella tabella 3.6 a pagina 39, relative rispettivamente ai *tweet* istituzionali e ai *tweet* non istituzionali, per ogni categoria viene evidenziata l'università con la percentuale più alta di *tweet*.

Nel caso istituzionale non vi è un ateneo predominante in molte categorie, mentre nel caso non istituzionale l'università Insubria risulta essere quella con percentuali più alte nelle categorie positive e le università Politecnico di Milano e Bicocca quelle con percentuali più alte nelle categorie

negative.

Istituzionali	**Positivi**					**Negativi**				
	Gen.	**Did.**	**Supp.**	**Sicur.**	**Amm.**	**Gen.**	**Did.**	**Supp.**	**Sicur.**	**Amm.**
Politecnico di Milano	5,1%	1,7%	4,0%	1,3%	1,0%	3,9%	2,0%	3,4%	1,3%	2,3%
Bicocca	1,3%	4,8%	4,2%	0,4%	2,4%	2,7%	1,3%	0,2%	0,3%	0,5%
Bologna	3,8%	1,4%	5,5%	0,9%	1,3%	1,4%	2,0%	2,6%	1,1%	1,5%
Brescia	3,1%	1,6%	8,7%	0,7%	0,4%	1,3%	2,5%	3,5%	0,5%	1,7%
Cafoscari	6,3%	3,2%	3,4%	1,3%	0,3%	4,2%	0,6%	2,0%	1,4%	0,8%
Iuav	5,1%	**6,0%**	1,1%	2,3%	1,0%	2,9%	1,1%	0,6%	**2,8%**	2,4%
Milano Statale	0,8%	2,5%	4,2%	0,8%	0,5%	1,3%	1,7%	3,5%	0,9%	0,8%
Parma	6,4%	1,4%	2,4%	1,2%	**3,0%**	**5,8%**	1,7%	1,7%	0,7%	2,4%
Pavia	4,2%	1,4%	2,8%	0,9%	0,0%	1,8%	2,4%	0,8%	2,3%	2,0%
Piemonte Orientale	1,3%	2,4%	**10,9%**	0,3%	0,4%	1,1%	1,0%	2,9%	0,8%	1,4%
Politecnico di Bari	4,6%	3,7%	7,2%	1,4%	2,0%	1,9%	1,4%	2,1%	0,8%	0,9%
Politecnico di Torino	2,0%	2,6%	3,6%	0,4%	0,5%	3,2%	1,1%	1,0%	2,0%	0,6%
Salento	2,0%	3,1%	5,1%	0,7%	1,3%	2,2%	**2,6%**	**4,1%**	0,9%	0,6%
Sassari	1,0%	1,8%	5,3%	0,7%	1,3%	4,5%	1,8%	2,6%	0,7%	2,4%
Udine	2,6%	3,1%	2,9%	1,1%	1,9%	3,1%	0,9%	1,3%	1,1%	1,8%
Verona	4,6%	1,7%	4,9%	1,8%	1,2%	1,7%	0,8%	1,2%	1,8%	**3,3%**
Sant'Anna	**6,4%**	2,4%	3,9%	3,0%	0,0%	3,3%	2,5%	1,7%	2,0%	1,5%
Sissa	2,7%	2,8%	2,4%	**3,6%**	1,3%	3,1%	2,2%	0,5%	1,6%	1,5%
SNS	3,0%	2,9%	5,5%	0,5%	1,2%	0,7%	0,3%	1,4%	0,9%	1,2%

Tabella 3.5: *Opinion Analysis* per i *tweet* istituzionali

Non Istituzionali	**Positivi**					**Negativi**				
	Gen.	**Did.**	**Supp.**	**Sicur.**	**Amm.**	**Gen.**	**Did.**	**Supp.**	**Sicur.**	**Amm.**
Politecnico di Milano	3,8%	1,1%	2,7%	1,5%	2,9%	3,1%	**4,3%**	4,1%	**2,9%**	2,0%
Bicocca	2,1%	0,8%	5,8%	1,3%	1,2%	**9,1%**	2,6%	**5,5%**	1,3%	1,0%
Bologna	4,8%	2,0%	2,1%	1,0%	1,0%	1,6%	2,0%	3,4%	1,9%	0,8%
Brescia	5,1%	3,7%	2,5%	4,1%	2,1%	1,8%	0,3%	0,7%	2,1%	2,2%
Cafoscari	5,7%	1,5%	3,0%	2,0%	1,1%	0,9%	1,6%	1,5%	1,4%	1,0%
Ferrara	4,6%	1,5%	3,8%	1,3%	**3,7%**	3,2%	2,7%	2,9%	1,4%	1,6%
Insubria	**11,5%**	**5,1%**	2,2%	**4,9%**	1,5%	4,5%	0,7%	0,5%	1,5%	**3,6%**
Iuav	3,3%	1,7%	3,9%	1,8%	0,6%	1,1%	0,9%	1,1%	1,2%	1,5%
Milano Statale	5,8%	4,9%	2,7%	2,4%	1,0%	1,6%	0,7%	2,5%	1,6%	1,0%
Parma	2,7%	2,1%	3,9%	0,4%	1,6%	4,6%	1,4%	0,8%	0,7%	2,3%
Pavia	3,3%	2,5%	6,9%	1,8%	1,5%	1,5%	1,2%	1,2%	1,3%	0,7%
Piemonte Orientale	2,6%	2,9%	2,2%	2,6%	1,3%	3,0%	0,9%	2,4%	2,3%	2,3%
Politecnico di Bari	2,7%	1,9%	6,4%	0,5%	1,1%	4,2%	2,2%	1,9%	2,5%	0,4%
Politecnico di Torino	2,3%	1,0%	3,4%	0,1%	0,7%	2,1%	1,7%	5,0%	2,8%	3,1%
Salento	2,8%	1,5%	2,5%	1,2%	0,5%	1,4%	2,0%	0,9%	1,7%	1,3%
Sassari	2,3%	2,6%	7,0%	2,5%	1,0%	2,2%	2,1%	0,8%	1,1%	1,5%
Trento	1,8%	1,3%	4,1%	0,5%	0,7%	3,9%	2,6%	1,6%	0,9%	1,5%
Udine	4,4%	2,0%	6,8%	1,9%	1,4%	1,3%	0,7%	1,3%	1,0%	0,8%
Verona	3,2%	1,7%	**8,2%**	1,7%	0,9%	2,3%	0,2%	0,7%	1,9%	1,5%
Sant'Anna	4,0%	1,8%	1,8%	2,4%	1,2%	4,0%	2,3%	4,3%	1,4%	1,4%
SNS	6,6%	0,8%	3,7%	0,9%	2,1%	2,3%	1,3%	3,2%	1,4%	1,4%

Tabella 3.6: *Opinion Analysis* per i *tweet* non istituzionali

3.2.5 Singola università: Politecnico di Milano

Sentiment Analysis e Analisi Temporale

Come ultima analisi, si è proceduto allo studio più dettagliata di una singola università, scegliendo come ateneo il Politecnico di Milano. Nella figura 3.12 è mostrata, sia per il caso istituzionale che per quello non istituzionale, la divisione dei tweet nelle categorie di Sentiment.

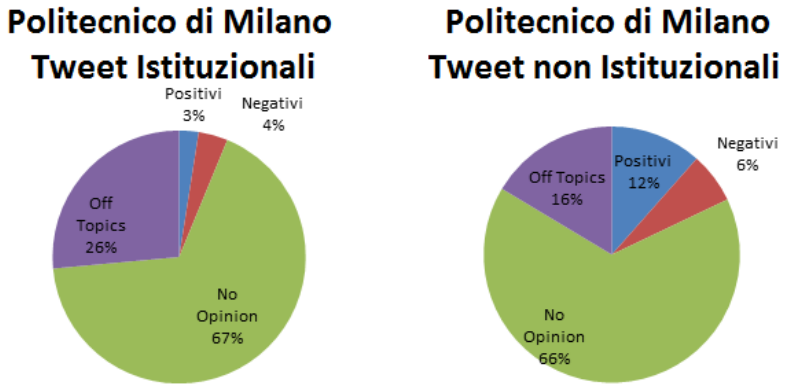


Figura 3.12: Distribuzione dei *tweet* per il Politecnico di Milano

Si è proceduto ad un'analisi dell'evoluzione temporale del Sentiment relativo al Politecnico di Milano nell'intervallo di tempo disponibile. La finestra temporale è stata scelta su frequenza settimanale. Nella figura 3.13 si può notate come il Sentiment Negativo e quello Positivo sembrino sostanzialmente accoppiati.

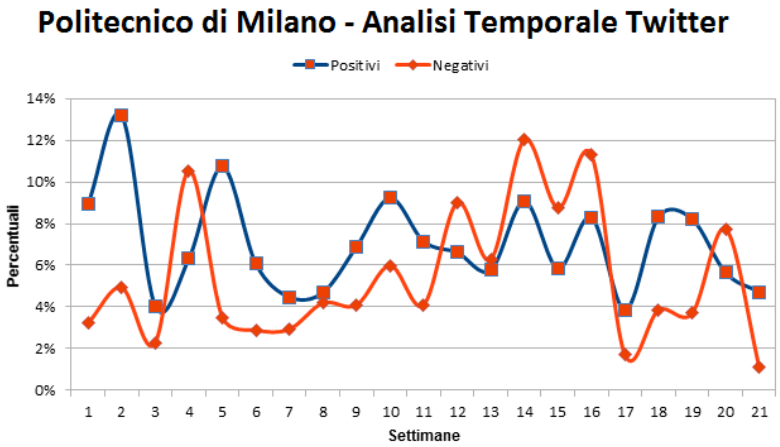


Figura 3.13: Analisi temporale dei *tweet* per il Politecnico di Milano

3.3 Facebook

La seconda analisi è stata effettuata sul *social network* Facebook studiando i diversi *post* relativi alle università. I *post* studiati hanno inizio dal primo Gennaio 2014 fino al 30 Giugno 2015. Una analisi descrittiva del dataset relativo a Facebook è mostrato nella tabella 3.7

Numero post totali	97595
Numero post istituzionali	5508
Numero post non istituzionali	92087
Numero di atenei analizzati	23
Numero medio di post per ateneo	∼4250

Tabella 3.7: Analisi descrittiva dataset Facebook

La divisione dei *post* nei vari atenei è raffigurata in figura 3.14. Si osserva come le numerosità siano significativamente diverse tra i singoli atenei e come la numerosità dei *post* istituzionali sia molto minore di quelli non istituzionali.

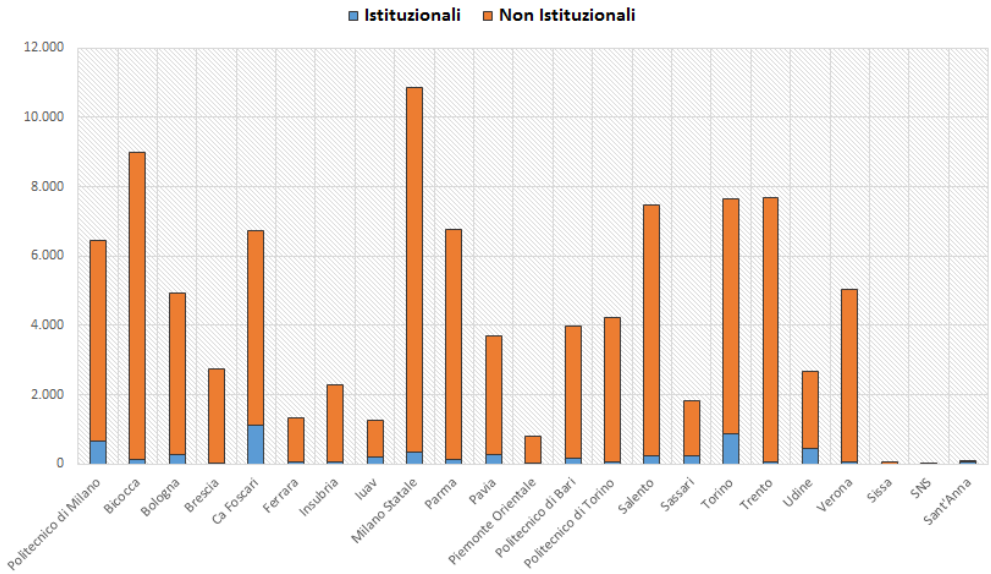


Figura 3.14: Distribuzione dei *post* per ateneo e per istituzionalità

3.3.1 Il training set e l'hand-coding

L'analisi è iniziata con la costruzione del *training set* e la seguente fase di *hand coding*. Il *training set*, formato da 1032 *post*, è stato costruito anche in questo caso cercando di bilanciare il numero di post nei diversi mesi e nei diversi atenei all'interno del *training set*.

Così come nel caso precedente di Twitter durante la procedura di *hand-coding* non si sono riscontrati *post* appartenenti alla categoria **Neutra**. Si è deciso dunque, anche in questo caso di eliminare questa categoria. La figura 3.15 mostra la distribuzione dei *post* nel *training set*.

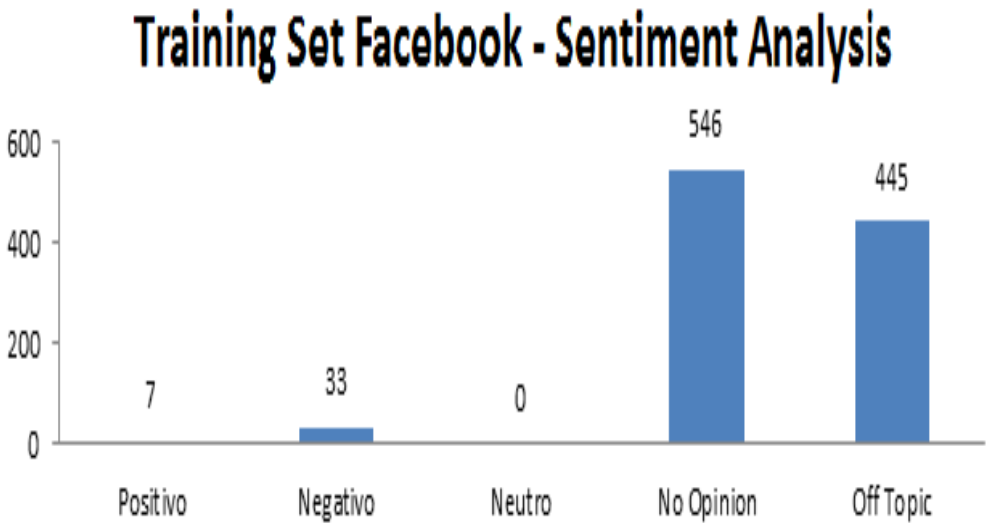


Figura 3.15: Numero *post* per le categorie di *sentiment analysis* nel *training set*

Durante la fase di *tagging* manuale per la *sentiment analysis*, si è proceduto in simultanea, al *tagging* per l'*opinion analysis*. I risultati ottenuti dall'*hand-coding* sono riportati in figura 3.16 nella pagina successiva. In questo caso non si sono riscontrati *post* relativi alle categorie positive di Didattica e Student Experience (C2), Sicurezza persone e asset fisici (C4) e Amministrazione e Governance (C5). Si è deciso di mantenerle nel modello poiché sono presenti alcuni *post* nelle categorie negative contrapposte.

Come nel caso di Twitter, la maggior parte dei *post* presenti nel *training set*, sono raccolti dalla categoria **Altro**, ovvero la categoria dell'*opinion analysis* che include le categorie di **No Opinion** e **Off Topics** della *senti-*

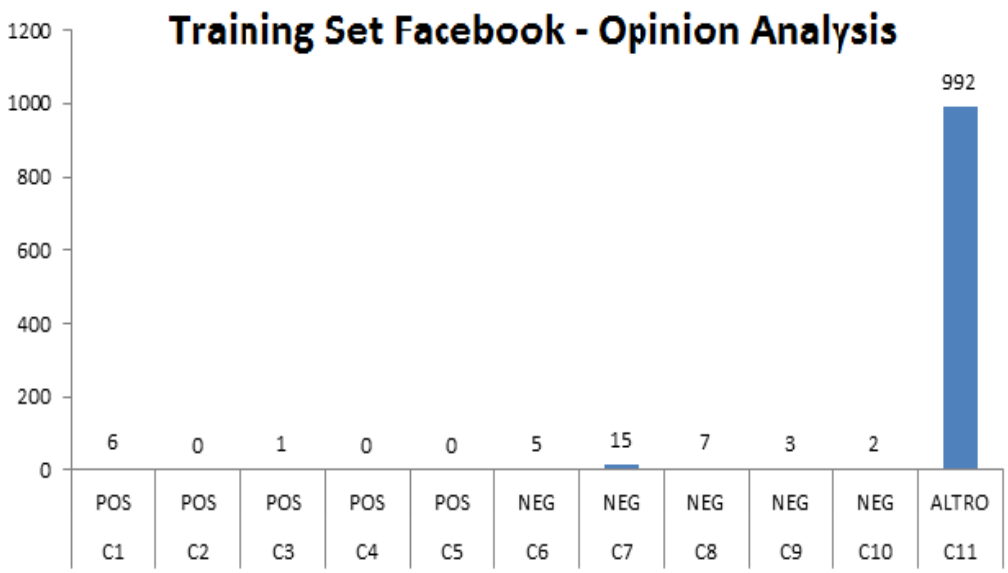


Figura 3.16: Numero *post* per le categorie di *opinion analysis* nel *training set*

ment analysis.

Tale risultato è in accordo con le informazioni a priori sulla modo di comunicare sui *social network* e sulle modalità di *download* dei *post*. Nonostante la bassa numerosità di alcune categorie di *opinion analysis*, si è deciso comunque di conservale distinte e non accorparle, per non perder l'informazione relativa ad ogni determinata categoria.

3.3.2 Elaborazione dei *post*

In questa fase vengono ripetuti gli stessi procedimenti seguiti per il dataset di Twitter. Poiché nel caso di Facebook, non vi è alcuna limitazione sul numero di caratteri di cui può essere composto un *post*, nello studio di questo *social network*, si ha un linguaggio significativamente diverso da quello di Twitter. Anche nel caso di Facebook non è stato possibile riconoscere le parole allungate arrivando dunque ad una perdita di informazioni. Così come in Twitter, si è utilizzato per la procedura di *stemming*, lo *stemmer* di Porter, presente nel *package tm* scaricabile in *R*.

3.3.3 Sistema Universitario

Utilizzando le configurazioni introdotte precedentemente, si è proceduto all'esecuzione della *sentiment analysis* e dell'*opinion analysis* dei *post* per l'intero sistema universitario.

Sentiment Analysis

La figura 3.17 mostra, per l'intero sistema universitario, sia nel caso di *post* istituzionali e non, la distribuzione degli stessi nelle quattro categorie analizzate.

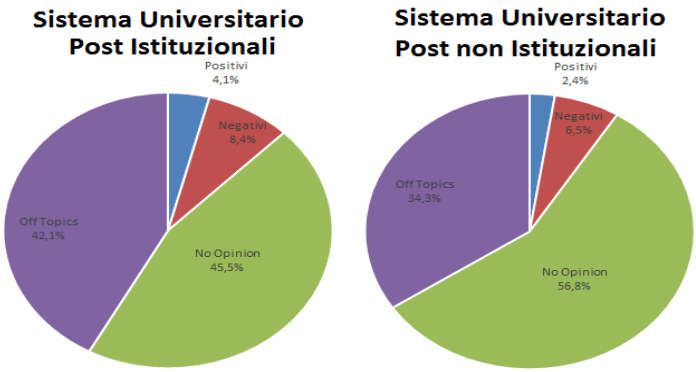


Figura 3.17: Distribuzione dei *post* per le categorie di *sentiment analysis* per l'intero sistema universitario

Si può osservare come, focalizzandosi sui *post* esprimenti giudizio, che le percentuali di *post* positivi e negativi sia maggiore nel caso di *post* istituzionali rispetto quelli non istituzionali. Come nel caso di Twitter, si è giustificato il fenomeno, supponendo che per assicurarsi che il messaggio giunga alle persone competenti, lo studente esprima il suo disappunto o la sua approvazione direttamente sulle pagine istituzionali dell'ateneo.
Chiaramente, occorre non dimenticare la diversa numerosità di *post*, come è visibile nella figura 3.18 nella pagina successiva

Opinion Analysis

In questa sezione mettiamo in evidenza la composizione dei giudizi negativi e positivi rispetto alle cinque sottocategorie dell'opinion analysis:

- Generico;
- Didattica & student experience;

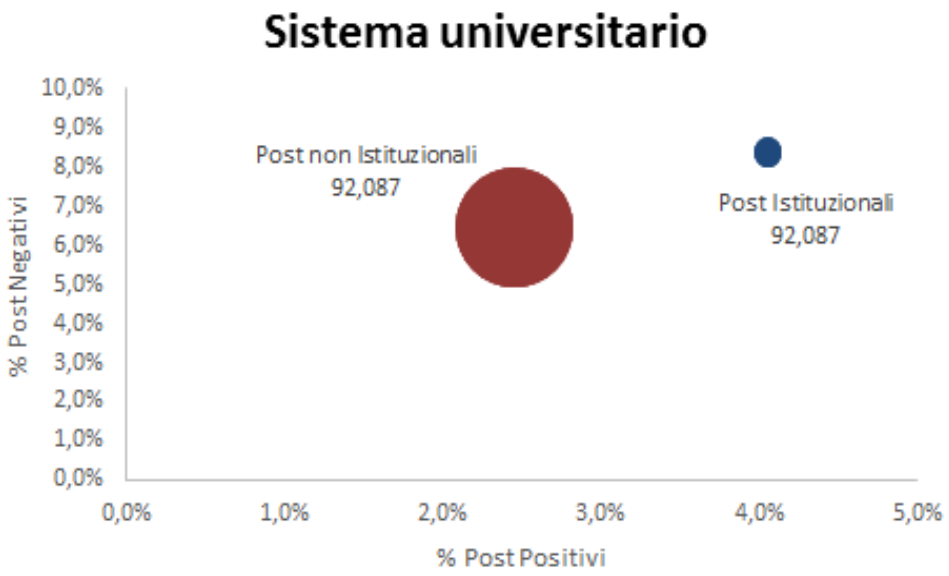


Figura 3.18: Confronto *post* con giudizio per l'intero sistema universitario per la *sentiment analysis*

- Supporto alla didattica;
- Sicurezza persone e asset fisici;
- Amministrazione e Governance.

Come si vede dalla figura 3.19 nella pagina seguente, nel caso di giudizi positivi, come ci aspettavamo, sono del tutto assenti le categorie Didattica & student experience (C2), Sicurezza persone e asset fisici (C4) e Amministrazione e Governance (C5). Gli aspetti più criticati, ovvero le categorie che presentano un maggior numero di *post* nel caso negativo, sono la **Didattica & student experience** e la **Supporto alla didattica** (che risultano dominanti sia nel caso istituzionale che in quello non istituzionale).

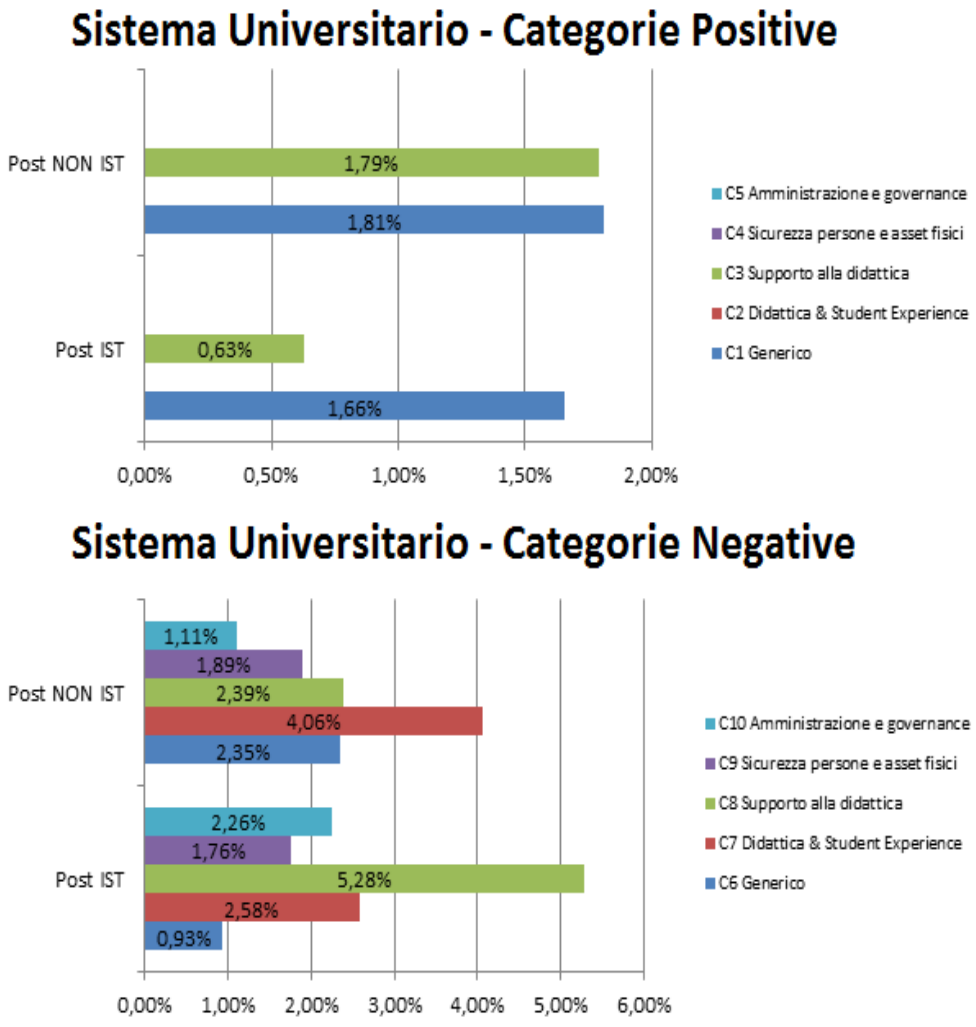


Figura 3.19: Confronto *post* con giudizio per l'intero sistema universitario per l'*opinion analysis*

3.3.4 Confronto tra gli atenei

Vogliamo adesso analizzare le differenze ricavate dalla *sentiment analylis* e dall'*opinion analysis* tra ogni singolo ateneo.

Ranking

Si vuole ora stilare una classifica che permetta di individuare gli atenei in cui gli studenti esprimono maggior disappunto. Si è costruito un indice sintetico dato dal rapporto tra la percentuale dei *post* negativi e la percentuale dei *post* con giudizio (quindi negativi e positivi). Ovvero l'indici di ranking sarà:

$$r_{neg} = \frac{\%postNegativi}{\%postNegativi + \%postPositivi}$$

Come si può notare dai risultati mostrati in figura 3.20, gli atenei occupano posizioni diverse a seconda se si considerano post istituzionali o meno. Nel primo caso si ha che l'ateneo con un *sentiment* negativo maggiore è il **Politecnico di Torino**, mentre nel caso dei *post* non istituzionali si ha che l'**Università degli studi dell'Insubria** presenta un r_{neg} più elevato.

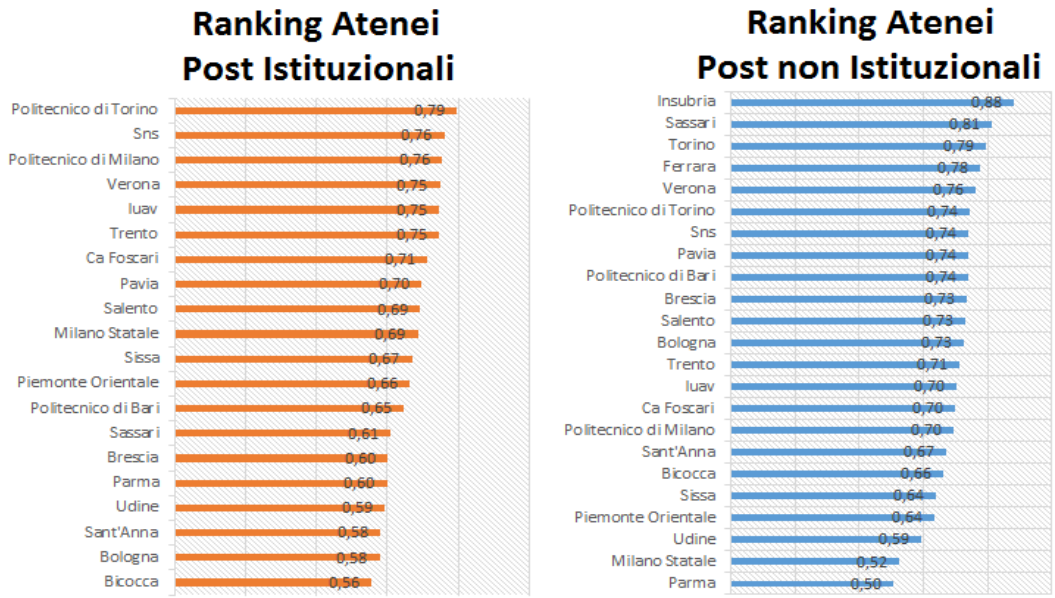


Figura 3.20: Ranking *post* con giudizio tra i diversi atenei per la *sentiment analysis*

Benchmarking

Si vuole adesso confrontare tra loro i diversi atenei in un grafico a bolle. Ciò permetterà di visualizzare graficamente anche il numero di *post* per ogni singola università, indice anche di affidabilità del risultato ottenuto (più è grande il campione di partenza e migliore sarà la stima ottenuta). Per evitare sovraffollamento e quindi rendere il grafico illeggibile, si sono selezionati gli atenei con almeno 200 *post* nel caso istituzionale e con almeno 5000 *post* nel caso non istituzionale.

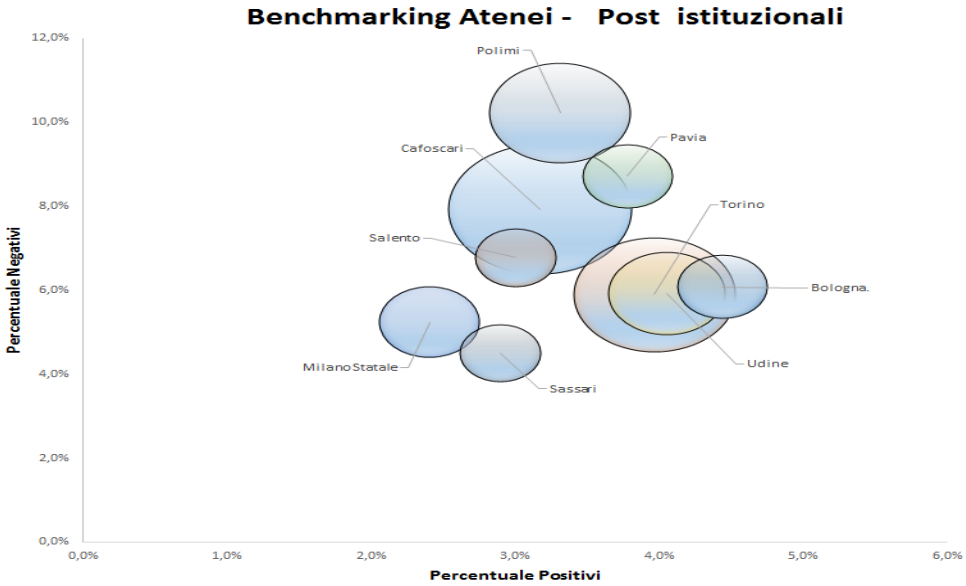


Figura 3.21: Benchmarking *post* istituzionali con giudizio tra i diversi atenei per la *sentiment analysis*

Dai grafici del benchmarking si possono ottenere dei rapidi confronti visivi tra le varie università. Nel caso istituzionale (figura 3.21), non si nota la presenza di gruppi ben distinti, solo le università di *Sassari* e *Milano Statale* sembrano leggermente staccate dalle altre. Anche nel caso non istituzionale (figura 3.22 nella pagina successiva) sembra esserci un unico gruppo con caratteristiche simili, da cui si distaccano le università di *Parma* e *Milano Statale*.

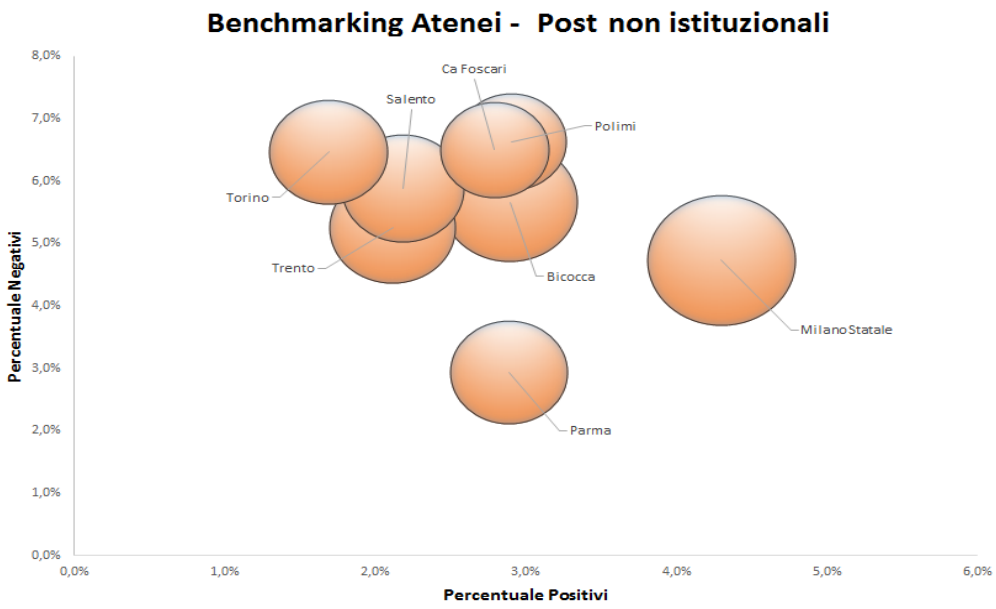


Figura 3.22: Benchmarking *post* non istituzionali con giudizio tra i diversi atenei per la *sentiment analysis*

Opinion Analysis

Per mostrare i risultati relativi all'*opinion analysis*, abbiamo scelto di utilizzare delle tabelle, in quanto permettono di confrontare tutte le categorie insieme, cosa che per leggibilità non sarebbe possibile facendo un grafico per ognuna delle cinque sottocategorie. In questo modo riduciamo tutto a solo due tabelle, dove le percentuali sono calcolate rispetto al totale dei post dell'ateneo.

Nella tabella 3.8 a fronte e nella tabella 3.9 a pagina 52, relative rispettivamente ai *post* istituzionali e ai *post* non istituzionali, per ogni categoria viene evidenziata l'università con la percentuale più alta di *post*.

Nel caso istituzionale l'ateneo Bicocca è predominante nelle categorie positive del supporto alla didattica e in quelle negative generiche e didattica e student experience. Da osservare anche il caso dell'università di Sassari che eccelle nelle categoria positiva generica mentre risulta la "peggiore" in quella di amministrazione e governance.

Nel caso non istituzionale la Scuola normale superiore (SNS) presenta le percentuali più elevate nelle categorie positive del supporto alla didattica e in quelle negative generiche, mentre il Politecnico di Bari, così come l'università di Sassari nel caso istituzionale, eccelle nelle categoria positiva generica mentre risulta la "peggiore" in quella di amministrazione e governance.

Istituzionali	Positivi		Negativi				
	Gen.	Supp.	Gen.	Did.	Supp.	Sicur.	Amm.
Polimi	1,8%	1,4%	1,7%	2,1%	**5,4%**	1,0%	0,6%
Bicocca	1,5%	**3,3%**	**3,4%**	**6,6%**	1,8%	1,9%	1,9%
Bologna	1,9%	2,3%	1,3%	5,1%	3,5%	0,8%	1,4%
Brescia	2,3%	1,4%	1,9%	5,1%	4,4%	1,5%	0,7%
Ca Foscari	**3,1%**	0,6%	1,7%	3,2%	2,8%	3,2%	1,1%
Ferrara	2,0%	1,4%	2,4%	3,1%	5,7%	1,7%	1,7%
Insubria	2,6%	1,7%	1,8%	1,7%	2,8%	5,9%	1,1%
Iuav	1,7%	0,0%	1,0%	5,4%	2,3%	1,2%	1,2%
MilanoStatale	2,6%	1,9%	1,1%	3,4%	3,7%	1,9%	0,9%
Parma	1,0%	0,0%	1,4%	4,3%	2,5%	**7,9%**	1,2%
Pavia	1,8%	1,2%	1,5%	1,8%	3,9%	1,5%	1,5%
P. Orien.	1,0%	2,4%	2,0%	2,9%	2,3%	1,4%	1,6%
Poliba	2,5%	0,0%	1,7%	3,4%	3,7%	1,6%	1,1%
Polito	1,1%	1,0%	0,9%	4,1%	2,6%	3,0%	0,6%
Salento	2,4%	0,0%	1,2%	3,8%	4,6%	0,5%	0,6%
Sassari	**3,1%**	2,7%	2,8%	4,1%	3,0%	2,6%	**2,2%**
Torino	2,3%	1,2%	2,0%	1,2%	4,6%	1,4%	1,2%
Trento	0,5%	3,0%	2,2%	6,0%	2,5%	2,9%	2,0%
Udine	3,0%	0,0%	2,4%	3,4%	1,6%	1,9%	0,0%
Verona	2,5%	0,0%	2,4%	2,8%	0,8%	6,1%	2,0%
Santanna	2,9%	0,0%	2,1%	2,9%	1,9%	2,3%	1,8%
Sissa	2,6%	0,0%	1,0%	2,8%	2,2%	2,4%	**2,2%**
SNS	1,1%	0,8%	1,9%	2,7%	4,9%	2,0%	1,4%

Tabella 3.8: *Opinion Analysis* per i *post* istituzionali

Non Istituzionali	Positivi		Negativi				
	Gen.	**Supp.**	**Gen.**	**Did.**	**Supp.**	**Sicur.**	**Amm.**
Politecnico di Milano	2,1%	0,7%	1,3%	4,2%	1,9%	**5,1%**	0,0%
Bicocca	2,8%	0,0%	1,3%	3,0%	3,5%	1,7%	1,2%
Bologna	2,0%	1,6%	1,2%	4,0%	3,9%	0,8%	2,1%
Brescia	1,9%	0,9%	1,7%	1,5%	2,5%	1,4%	1,6%
Ca Foscari	1,3%	0,8%	1,7%	2,4%	1,7%	1,4%	0,9%
Ferrara	2,3%	0,0%	1,4%	3,1%	**6,6%**	1,3%	0,9%
Insubria	1,9%	0,7%	1,8%	4,7%	3,6%	1,3%	1,7%
Iuav	1,4%	1,3%	3,1%	5,9%	3,7%	2,5%	0,8%
Milano Statale	2,9%	1,9%	1,5%	2,5%	3,7%	0,9%	0,8%
Parma	2,1%	1,8%	0,9%	2,1%	5,3%	0,8%	1,2%
Pavia	1,2%	0,9%	2,0%	5,0%	2,1%	3,4%	1,3%
Piemonte Orientale	1,5%	2,5%	3,0%	3,9%	1,0%	1,2%	1,3%
Politecnico di Bari	**3,0%**	0,0%	2,2%	4,5%	2,3%	2,0%	**4,0%**
Politecnico di Torino	2,1%	1,4%	2,2%	**7,9%**	1,5%	2,7%	2,5%
Salento	1,8%	1,9%	2,0%	4,8%	1,5%	2,3%	1,3%
Sassari	1,7%	1,2%	1,6%	3,0%	3,1%	1,0%	1,0%
Torino	2,2%	1,9%	1,2%	3,5%	3,5%	0,0%	0,9%
Trento	1,4%	2,0%	1,6%	3,1%	3,3%	2,4%	0,9%
Udine	2,6%	1,6%	2,1%	3,6%	2,7%	2,5%	1,5%
Verona	2,9%	0,8%	1,6%	3,5%	3,7%	0,8%	1,3%
Santanna	1,8%	2,8%	2,0%	2,5%	2,8%	1,9%	0,8%
Sissa	1,5%	1,2%	1,5%	3,7%	4,0%	1,9%	1,0%
SNS	2,1%	**3,4%**	**3,2%**	3,2%	1,6%	4,8%	0,0%

Tabella 3.9: *Opinion Analysis* per i *post* non istituzionali

3.3.5 Singola università: Politecnico di Milano

Sentiment Analysis e Analisi Temporale

Come ultima analisi, si è proceduto allo studio più dettagliata di una singola università, scegliendo come ateneo il Politecnico di Milano. Nella figura 3.23 è mostrata, sia per il caso istituzionale che per quello non istituzionale, la divisione dei post nelle categorie di Sentiment.

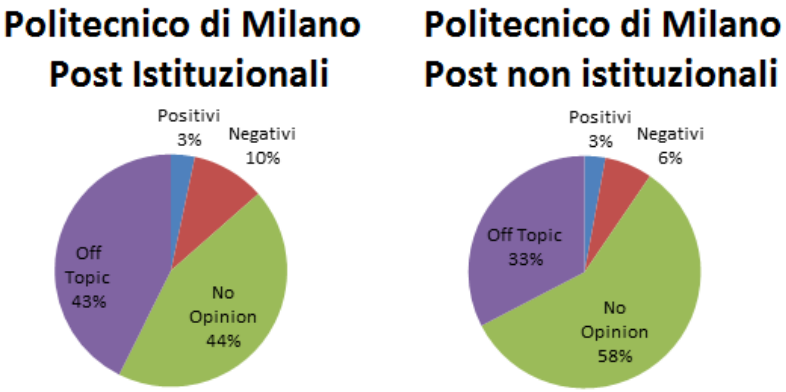


Figura 3.23: Distribuzione dei *post* per il Politecnico di Milano

Si è proceduto ad un'analisi dell'evoluzione temporale del Sentiment relativo al Politecnico di Milano nell'intervallo di tempo disponibile. La finestra temporale è stata scelta su frequenza settimanale, i risultati sono mostrati nelle figura 3.24 nella pagina successiva e 3.25 nella pagina seguente Si può notare come sia per il 2014 che per il 2015, il sentiment Positivo e quello Negativo sembrano in qualche modo accoppiati. Ciò significa che durante l'anno ciò che più varia è la proporzione tra sentiment con giudizio e sentiment che ne è privo.

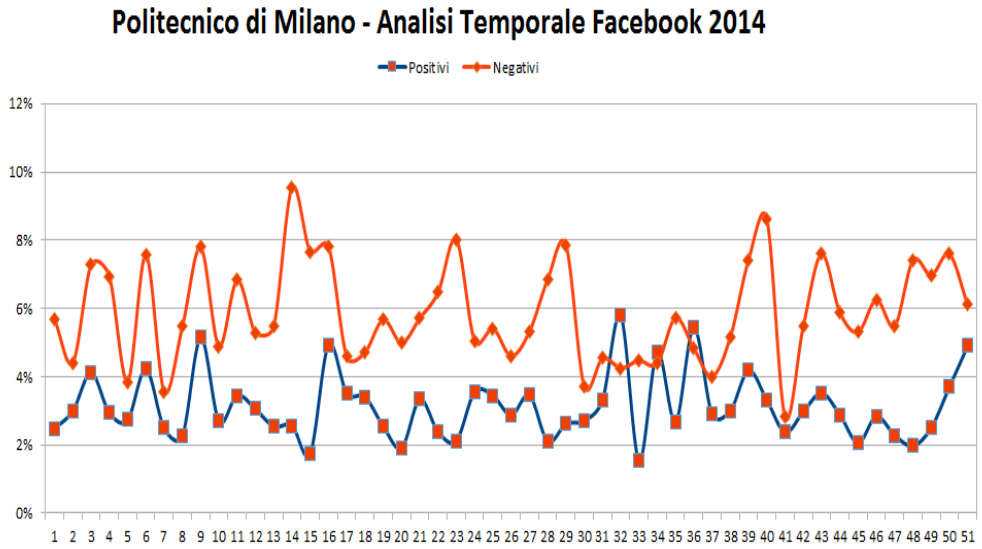


Figura 3.24: Analisi temporale dei *tweet* del 2014 per il Politecnico di Milano

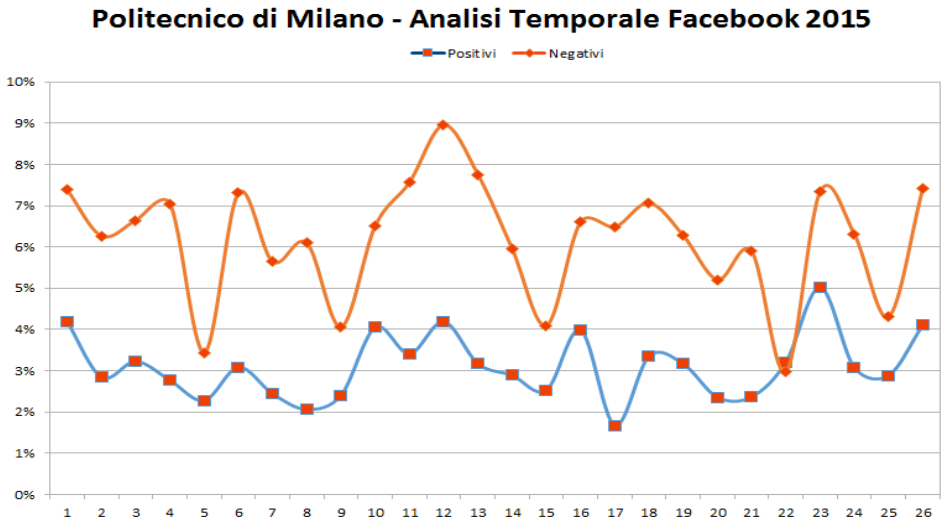


Figura 3.25: Analisi temporale dei *tweet* del 2015 per il Politecnico di Milano

3.4 Regressione Logistica

In ultima analisi, si è cercato di individuare, dati dei possibili regressori, un modello matematico che permettesse di spiegare le variazioni di sentiment in funzione di alcune caratteristiche di ogni singolo ateneo. La regressione logistica è un caso particolare di modello lineare generalizzato. Si tratta di un modello di regressione applicato nei casi in cui la variabile dipendente y sia di tipo dicotomico riconducibile quindi ai valori 0 o 1. Il modello viene descritto da:

$$logit(p) = \ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k = X\beta \tag{3.2}$$

In questa analisi esplorativa, si sono considerate come unità statistiche alla base dei modelli di regressione logistica esaminati, i singoli *post* o *tweet*, trascurando per semplicità le possibili dipendenze tra i messaggi scritti dai vari utenti e la presenza di eventuali periodicità e/o dinamiche temporali.

Inoltre nel nostro caso si considereranno i soli modelli additivi, ovvero quei modelli in cui vengono trascurati i termini di interazione tra i vari regressori. Le regressioni logistiche effettuate considerano come variabile bernoulliana di risposta il sentiment dei *post* o dei *tweet*, in modo da poter analizzare le tipologie:

- Sentiment Negativo contro la restante parte;
- Sentiment Positivo contro la restante parte;
- Sentiment Negativo di una particolare categoria contro la restante parte;
- Sentiment Positivo di una particolare categoria contro la restante parte;

Inoltre soffermandosi solamente sui messaggi contenenti un giudizio (ovvero i soli *tweet* o *post* positivi o negativi), si sono investigate:

- Sentiment Negativo contro Sentiment Positivo;
- Sentiment Negativo di una particolare categoria contro i restanti messaggi con giudizio;

- Sentiment Positivo di una particolare categoria contro i restanti messaggi con giudizio;

3.4.1 Regressori

I regressori che sono stati analizzati nell'eseguire le regressioni logistiche sono:

- **Dimensione**: Variabile numerica corrispondente alla dimensione in numero di studenti dell'ateneo;
- **Propensione**: Variabile numerica rappresentativa della propensione digitale degli studenti dell'ateneo, calcolata come numero di *post* o *tweet* procapite: $propensione = \frac{\#messaggi}{\#studenti}$;
- **Locazione Geografica**: Variabile categorica (Nord o Centro-Sud) corrispondente all'area di appartenenza dell'ateneo;
- **Tipologia**: Variabile categorica (Tecnica o Generalista) corrispondente alla tipologia di appartenenza dell'ateneo;
- **Proattività**: Variabile numerica rappresentativa dell'attività istituzionale dell'ateneo sui canali social, calcolata come il numero di *post* o *tweet* certificati istituzionali (ovvero scritti dal profili autorizzati dall'ateneo).
- **Rank**: Variabile booleana (Si o No) rappresentativa della presenza o meno dell'ateneo nella prime 400 posizioni del QS ranking.

Nelle tabelle 3.10 nella pagina successiva e 3.11 a pagina 58 si riportano i valori su cui si sono effettuate le regressioni.

Si sono quindi analizzati 40 diversi modelli: 23 nello studio di Twitter e 17 in quello di Facebook, poichè dallo studio dell'Opinion Analysis nel caso di Facebook sono risultate assente tre categorie positive (Didattica e Student Experience, Sicurezza persone e asset fisici e Amministrazione e Governance) invece presenti nel caso di Twitter.

Data l'elevata numerosità dei test effettuati, si è applicata una correzione di Bonferroni ai p-value ottenuti come risultati dei molteplici modelli analizzati. Si è così controllata la probabilità di effettuare false scoperte correggendo i p-value: $pvalue_{bonf} = pvalue \times n$.

Ateneo	Dim.	Prop.	Locazione	Tipologia	Proat.	Rank
Polimi	41.613	0,155	Nord	Tecnica	144	Si
Unimib	32.407	0,277	Nord	Generalista	67	No
Unibo	84.215	0,059	Centro-Sud	Generalista	138	Si
Unibs	14.412	0,191	Nord	Generalista	79	No
CaFoscari	19.584	0,344	Nord	Generalista	266	No
Unife	15.865	0,084	Centro-Sud	Generalista	67	No
Uninsubria	8.854	0,260	Nord	Generalista	93	No
Iuav	4.607	0,275	Nord	Generalista	51	No
LaStatale	61.473	0,177	Nord	Generalista	132	Si
Unipr	25.371	0,267	Centro-Sud	Generalista	127	No
Unipv	21.829	0,169	Nord	Generalista	212	No
UniAvogadro	10.314	0,077	Nord	Generalista	9	No
Poliba	10.557	0,378	Centro-Sud	Tecnica	105	No
Polito	33.242	0,127	Nord	Tecnica	55	Si
Unisalento	19.025	0,392	Centro-Sud	Generalista	144	No
Uniss	13.116	0,139	Centro-Sud	Generalista	40	No
Torino	64.317	0,119	Nord	Generalista	285	No
Unitn	16.854	0,455	Nord	Generalista	71	No
Uniud	15.514	0,173	Nord	Generalista	242	No
Univerona	22.603	0,224	Nord	Generalista	63	No
Sissa	70	1,000	Nord	Generalista	14	No
ScuolaSantAnna	280	0,250	Centro-Sud	Generalista	25	No
ScuolaNormale	66	0,258	Centro-Sud	Generalista	14	No

Tabella 3.10: Valori dei vari regressori per ogni ateneo nello studio di Facebook

I risultati ottenuti dall'analisi dei modelli sono riportati nella tabelle 4.1 a pagina 65 e 4.2 a pagina 66 in Appendice A (i $pvalue_{bonf}$ sono stati limitati ad 1). Si è evidenziato per ogni singolo regressore, il p-value ottenuto come output del modello e un indice sintetico ($+1$,-1 e 0) per indicarne l'apporto positivo,negativo o nullo alla variazione di probabilità. Le prime tre righe di ogni tabella analizzano il sentiment dei due social network senza scendere nel dettaglio della singola categoria di opinion.

Nel caso di Twitter si ha che l'appartenenza ad una tipologia tecnica e all'aumentare della dimensione dell'ateneo, aumenti la probabilità che il *tweet* non sia positivo (no opinion/off topic o negativo). Viceversa essere nelle prima 400 posizioni del QS ranking e all'aumentare della proattività, aumenti la probabilità che il *tweet* sia positivo. Per quanto riguarda invece

Ateneo	Dim.	Prop.	Locazione	Tipologia	Proat.	Rank
CaFoscari	19.584	0,136	Nord	Generalista	473	No
Iuav	4.607	0,217	Nord	Generalista	45	No
LaStatale	61.473	0,121	Nord	Generalista	457	Si
Poliba	10.557	0,067	Centro-Sud	Tecnica	102	No
Polimi	41.613	0,230	Nord	Tecnica	353	Si
Polito	33.242	0,084	Nord	Tecnica	96	Si
ScuolaNormale	66	35,636	Centro-Sud	Generalista	437	No
ScuolaSantAnna	280	3,400	Centro-Sud	Generalista	262	No
Sissa	70	0,614	Nord	Generalista	42	No
UniAvogadro	10.314	0,056	Nord	Generalista	223	No
Unibo	84.215	0,111	Centro-Sud	Generalista	1.651	Si
Unibs	14.412	0,075	Nord	Generalista	142	No
Unife	15.865	0,058	Centro-Sud	Generalista	0	No
Unimib	32.407	0,153	Nord	Generalista	289	No
Uninsubria	8.854	0,066	Nord	Generalista	0	No
Unipr	25.371	0,068	Centro-Sud	Generalista	37	No
Unipv	21.829	0,136	Nord	Generalista	585	No
Uniss	13.116	0,052	Centro-Sud	Generalista	75	No
Unitn	16.854	0,068	Nord	Generalista	0	No
Uniud	15.514	0,183	Nord	Generalista	554	No
Univerona	22.603	0,063	Nord	Generalista	382	No
Unisalento	19.025	0,032	Centro-Sud	Generalista	101	No

Tabella 3.11: Valori dei vari regressori per ogni ateneo nello studio di Twitter

la localizzazione al nord, questa sembra incidere negativamente sul sentiment dei *tweet*. Infine la propensione non sembra essere significativa nella variazione del sentiment.

I risultati ottenuti nel caso di Facebook invece mostrano come la dimensione, la localizzazione e appartenenza al QS ranking dell'ateneo non abbiamo alcuna influenza sulla variazione del sentiment. In accordo con Twitter si ha che la Proattività dell'ateneo aumenti la probabilità che il *post* abbia un sentiment positivo. La propensione in questo caso risulta aumentare la probabilità che il *post* non sia negativo (no opinion/off topic o positivo). Infine la tipologia tecnica risulta essere una discriminante a favore dei *post* che non contengono giudizi.

I modelli di regressione logistica che analizzano l'opinion analysis risultano essere meno leggibili e di difficile interpretazione. In linea generale non

si sono presenti forti incoerenze tra i due *social network* (regressori con contributo opposto), anche se alcuni regressori che non risultano significativi per un *social network*, possono esserlo nell'altro.

Capitolo 4

Conclusioni

Lo scopo di questo elaborato è stato lo studio dei due principali social network tramite Sentiment e Opinion Analysis. Il metodo scelto per effettuare le varie analisi è stato quello di Hopkins e King, uno degli attuali metodi più utilizzati in letteratura.

Le prime problematiche incontrate hanno riguardo la scelta delle categorie di Sentiment e Opinion e la successiva fase di *Hand-Coding*. Infatti, nonostante la scelta delle possibili classi di appartenenza dei diversi *tweet* o *post* sia stata fatta in modo da produrre insiemi mutuamente esclusi e che coprissero il più possibile tutto lo spettro delle varie possibilità, in fase di *tagging* manuale più volte si è incorsi in dubbi sull'appartenenza del singolo testo. Considerate queste criticità si è deciso di eseguire la procedura di *Hand-Coding* due volte, la prima per etichettare tutti i testi del *training-set*, mentre la seconda volta per affinare la scelta nei casi di dubbia appartenenza.
Nonostante la presenza di più *hand-coders* e la ripetizione della procedura sono comunque presenti degli errori inevitabili nella fase di addestramento della l'algoritmo. Il problema è anche legato alla scelta del tema analizzato. In questo caso l'argomento studiato non produce dei giudizi facilmente riconoscibili, come invece potrebbe essere ad esempio un referendum o un dibattito su un *trend topic* del momento. In questi casi infatti, pur essendo presenti dei testi privi di giudizio e fuori tema, nel caso di messaggi portatori di *sentiment*, si riscontra un accuratezza nella scelta della classe di appartenenza migliore.
Due ulteriori difficoltà che si sono incontrare nello svolgimento delle analisi sono di carattere computazionale. Innanzitutto si sottolinea la problema-

tica riguardo alla gestione e al preprocessamento dei testi in italiano. Nonostante i più recenti pacchetti fruibili in R permettano la gestione della lingua italiana, l'accuratezza dei risultati ottenuti non è ancora equiparabile alla bontà dei risultati in lingua inglese. Infine occorre rimarcare il notevole tempo computazionale richiesto per alcune delle analisi sui dataset più numerosi. In alcune occasioni il tempo occorso a superato la decina di ore.

Fatte salve le problematiche presentate, si è proceduto all'applicazione delle Sentiment e Opinion Analysis al *case study* degli atenei italiani. Si sono studiati quelli che attualmente risultano essere i due principali social network: *Facebook* e *Twitter*. Si sono voluti analizzare separatamente i *post* e i *tweet* istituzionali (ovvero quei messaggi pubblicati da soggetti autorizzati dall'ateneo) e i non istituzionali. In tutti e quattro i casi analizzati, come prevedibile, le classi più numerose sono risulta essere le *No-Opinion* e *Off-Topics*. Sia per *Facebook* che per *Twitter* si è riscontrata una diminuizione dei messaggi contenenti giudizio (siano essi positivi o negativi) nel caso non istituzionale.
Il confronto tra i due *social network* ha invece evidenziato una maggiore propensione ad esprimere giudizio su *Twitter*. Ciò sembra essere legato al diverso linguaggio presente sul social di Palo Alto, infatti qui si è forse più inclini alla condivisione di vicende personali generiche e non solo legate dunque ad esperienze positive o negative.

In ultima analisi si è proceduto allo studio delle regressioni logistiche che non ha evidenziato dei particolari legami per quanto riguarda i regressori analizzati tra i due diversi *social network*, eccenzion fatta per la proattività dell'ateneo (ovvero il numero di *post* o *tweet* scritti direttamente dall'ateneo) che è risultata essere significativa per un immagine migliore e positiva dell'università stessa. Dunque da quest'ultimo risultato si evince come università maggiormente attive sul web, in grado di coinvolgere periodicamente tramite i canali *touchpoint* i propri studenti, riescono ad fornire una reputazione migliore.

Come ultimo punto d'attenzione, si vuole concludere questo elaborato con uno sguardo ai possibili sviluppi futuri. Le attuale tendenze di *social sharing* sono sempre più affini al mondo *mobile*, infatti si stima che la

percentuale (in continua crescità) di utenti che interagiscono tramite l'*app* del *social network* sono circa il 60% del totale. Ciò rende possibile, nella maggior parte dei casi, la geolocalizzazione del messaggio, ovvero poter individuare da quale zona geografica è stato inviato. Si potrebbero così condurre delle analisi geostatistiche sul possibile andamento del *sentiment* sia nel tempo che nello spazio.

Un'ulteriore miglioria alle analisi effettuate potrebbe esssere l'applicazione delle più moderne *social network analysis*, in grado di cogliere i legami tra i vari utenti della rete, analizzando dunque per ognuno di essi la *social influence*, individuando i possibili *advocates* o *protesters* della reputazione dell'oggetto in esame.

Appendice A: Risultati regressioni Logistiche

Modello	Dim.	Prop.	Loc.Nord	Tip.Tecnica	Proat.	RankSi
FB1 PosVSAll	0 (1,000)	0 (1,000)	0 (1,000)	-1 (0,000)	1 (0,001)	0 (1,000)
FB2 NegVSAll	0 (1,000)	-1 (0,000)	0 (1,000)	-1 (0,000)	0 (1,000)	0 (1,000)
FB3 PosVSNeg	0 (1,000)	1 (0,000)	0 (1,000)	0 (1,000)	1 (0,000)	0 (1,000)
FB4 C1VSAll	0 (1,000)	0 (1,000)	-1 (0,000)	0 (1,000)	0 (1,000)	0 (1,000)
FB5 C3VSAll	0 (1,000)	-1 (0,000)	1 (0,000)	-1 (0,000)	1 (0,000)	0 (1,000)
FB6 C1VSOth	0 (1,000)	0 (1,000)	-1 (0,000)	0 (1,000)	0 (1,000)	0 (1,000)
FB7 C3VSOth	0 (1,000)	-1 (0,000)	1 (0,000)	-1 (0,000)	1 (0,000)	0 (1,000)
FB8 C6VSAll	0 (1,000)	0 (0,158)	1 (0,000)	0 (1,000)	0 (1,000)	1 (0,000)
FB9 C7VSAll	1 (0,011)	0 (1,000)	1 (0,000)	1 (0,000)	-1 (0,000)	-1 (0,000)
FB10 C8VSAll	-1 (0,007)	0 (0,379)	-1 (0,000)	1 (0,000)	0 (0,428)	0 (0,068)
FB11 C9VSAll	0 (1,000)	0 (1,000)	1 (0,000)	-1 (0,000)	-1 (0,000)	0 (1,000)
FB12 C10VSAll	0 (0,100)	0 (0,071)	-1 (0,000)	1 (0,000)	1 (0,000)	-1 (0,000)
FB13 C6VSOth	0 (1,000)	1 (0,032)	1 (0,001)	0 (0,967)	0 (1,000)	1 (0,000)
FB14 C7VSOth	1 (0,000)	0 (1,000)	1 (0,000)	1 (0,000)	-1 (0,000)	-1 (0,000)
FB15 C8VSOth	-1 (0,001)	1 (0,004)	-1 (0,000)	0 (1,000)	1 (0,000)	1 (0,000)
FB16 C9VSOth	0 (1,000)	0 (1,000)	1 (0,000)	-1 (0,000)	-1 (0,000)	0 (0,136)
FB17 C10VSOth	1 (0,023)	0 (0,060)	-1 (0,000)	1 (0,000)	1 (0,000)	-1 (0,000)

Tabella 4.1: Risultati regressioni logistiche nello studio di Facebook: VSOth altri post con giudizio, VSAll contro tutti gli altri post

Modello	Dim.	Prop.	Loc.Nord	Tip.Tecnica	Proat.	RankSi
TW1	-1	0	0	-1	1	1
PosVSAll	(0,000)	(1,000)	(1,000)	(0,000)	(0,000)	(0,000)
TW2	0	0	1	0	-1	-1
NegVSAll	(1,000)	(1,000)	(0,001)	(1,000)	(0,000)	(0,000)
TW3	-1	0	-1	-1	1	1
PosVSNeg	(0,000)	(1,000)	(0,004)	(0,000)	(0,000)	(0,000)
TW4	1	0	1	0	-1	-1
C1VSAll	(0,000)	(0,813)	(0,000)	(0,709)	(0,017)	(0,000)
TW5	1	-1	1	0	0	-1
C2VSAll	(0,000)	(0,001)	(0,000)	(1,000)	(0,841)	(0,000)
TW6	1	-1	1	0	0	0
C3VSAll	(0,000)	(0,000)	(0,000)	(0,146)	(1,000)	(0,060)
TW7	1	0	0	0	0	-1
C4VSAll	(0,000)	(0,082)	(0,055)	(1,000)	(1,000)	(0,000)
TW8	1	-1	0	0	0	-1
C5VSAll	(0,000)	(0,000)	(1,000)	(1,000)	(1,000)	(0,000)
TW9	0	0	0	0	0	-1
C6VSAll	(1,000)	(1,000)	(0,138)	(0,370)	(1,000)	(0,000)
TW10	1	0	1	0	0	0
C7VSAll	(0,001)	(0,269)	(0,004)	(1,000)	(1,000)	(0,294)
TW11	0	0	1	0	0	0
C8VSAll	(1,000)	(0,187)	(0,000)	(0,130)	(1,000)	(1,000)
TW12	1	0	0	0	0	-1
C9VSAll	(0,001)	(1,000)	(1,000)	(1,000)	(1,000)	(0,000)
TW13	0	-1	0	0	0	-1
C10VSAll	(0,374)	(0,000)	(1,000)	(1,000)	(1,000)	(0,000)
TW14	0	1	0	0	0	0
C1VSOth	(0,364)	(0,001)	(1,000)	(1,000)	(1,000)	(0,221)
TW15	0	0	-1	0	0	0
C2VSOth	(1,000)	(1,000)	(0,000)	(1,000)	(1,000)	(0,392)
TW16	-1	1	-1	0	0	1
C3VSOth	(0,000)	(0,000)	(0,000)	(1,000)	(0,513)	(0,000)
TW17	1	-1	1	0	0	0
C4VSOth	(0,003)	(0,000)	(0,019)	(1,000)	(1,000)	(0,136)
TW18	0	-1	0	0	0	0
C5VSOth	(1,000)	(0,000)	(1,000)	(1,000)	(1,000)	(1,000)
TW19	-1	1	-1	0	0	1
C6VSOth	(0,000)	(0,000)	(0,000)	(0,096)	(1,000)	(0,000)
TW20	-1	1	-1	0	0	1
C7VSOth	(0,033)	(0,026)	(0,000)	(1,000)	(1,000)	(0,000)
TW21	-1	1	-1	0	0	1
C8VSOth	(0,000)	(0,000)	(0,000)	(0,907)	(1,000)	(0,000)
TW22	0	-1	0	0	0	0
C9VSOth	(1,000)	(0,009)	(1,000)	(1,000)	(1,000)	(1,000)
TW23	0	0	0	0	0	1
C10VSOth	(1,000)	(0,171)	(1,000)	(1,000)	(1,000)	(0,024)

Tabella 4.2: Risultati regressioni logistiche nello studio di Twitter: VSOth altri post con giudizio, VSAll contro tutti gli altri post

Appendice B: Stopwords

a	abbia	abbiamo	abbiano	abbiate	ad
agl	agli	ai	al	all	alla
alle	allo	anche	anziche	anziché	anzichè
avemmo	avendo	avesse	avessero	avessi	avessimo
aveste	avesti	avete	aveva	avevamo	avevano
avevate	avevi	avevo	avrà	avrai	avranno
avrebbe	avrebbero	avrei	avremmo	avremo	avreste
avresti	avrete	avrò	avuta	avute	avuti
avuto	b	c	che	chi	chissà
ci	ció	ciò	cmq	coi	col
come	comunque	con	contro	cose	così
cui	d	da	dà	dagl	dagli
dai	dal	dall	dalla	dalle	dallo
degl	degli	dei	del	dell	della
delle	dello	di	dov	dove	è
e	é	ebbe	ebbero	ebbi	ecc
ed	era	erano	eravamo	eravate	eri
ero	essendo	etc	f	fa	faccia
facciamo	facciano	facciate	faccio	facemmo	facendo
facesse	facessero	facessi	facessimo	faceste	facesti
faceva	facevamo	facevano	facevate	facevi	facevo
fai	fanno	farà	farai	faranno	farebbe
farebbero	farei	faremmo	faremo	fareste	faresti
farete	farò	fece	fecero	feci	fosse
fossero	fossi	fossimo	foste	fosti	fu
fui	fummo	furono	g	già	gli
h	ha	hai	hanno	ho	i

il	in	io	j	k	l
la	le	lei	li	lo	loro
lui	m	ma	mi	mia	mie
miei	mio	n	ne	né	nè
negl	negli	nei	nel	nell	nella
nelle	nello	noi	non	nostra	nostre
nostri	nostro	o	ogni	p	per
perche	perché	perchè	però	più	po
pò	poi	può	q	qual	quale
quali	quando	quanta	quante	quanti	quanto
quell	quella	quelle	quelli	quello	quest
questa	queste	questi	questo	r	s
sa	sarà	sarai	saranno	sarebbe	sarebbero
sarei	saremmo	saremo	sareste	saresti	sarete
sarò	se	sei	si	sia	siamo
siano	siate	siete	sono	st	sta
stai	stando	stanno	starà	starai	staranno
starebbe	starebbero	starei	staremmo	staremo	stareste
staresti	starete	starò	stava	stavamo	stavano
stavate	stavi	stavo	stemmo	stesse	stessero
stessi	stessimo	steste	stesti	stette	stettero
stetti	stia	stiamo	stiano	stiate	sto
su	sua	sue	sugl	sugli	sui
sul	sull	sulla	sulle	sullo	suo
suoi	t	ti	tra	tu	tua
tue	tuo	tuoi	tutti	tutto	u
un	una	uno	v	vabbè	vi
w	xché	xchè	x	xche	xkè
y	via	voi	vostra	vostre	vostri
vostro	z				

Bibliografia

Amati, G., Bianchi, M. & Marcone, G. (2014), 'Sentiment estimation on twitter', *Proceedings of the 5th Italian Information RetrievalWorkshop* pp. 39–50.

Blei, D. M. (2012), 'Surveying a suite of algorithms that offer a solution to managing large document archives', *communications of the acm* .

Cacco, F. (2012-2013), 'Metodo hopkins-king per la sentiment analysis: una valutazione basata sui tweets della campagna elettorale', *Università degli studi di Padova* .

Ceron, A., Curini, L. & Iacus, S. (2014), *Social media and sentiment analysis. l'evoluzione dei fenomeni sociali attraverso la rete*, Springer Editore.

Cohen, J. (1960), 'A coefficient of agreement for nominal scales', *Educational and Psychological Measurement* pp. 37–46.

Feinerer, I. (2014), *Pacchetto tm: text mining package.*

Fleiss, J. L. (1971), 'Measuring nominal scale agreement among many raters', *Bulletin* pp. 378–382.

Gagolewski, M. & Tartanus, B. (2014), *Pacchetto stringi : character string processing facilities.*

Gentry, J. (2015), *Pacchetto TwitteR: R based twitter client.*

Girotto, A. D. (2012-2013), 'Sentiment analysis con il metodo hopkings-king: studio dei commenti ad un post di un blog', *Università degli studi di Padova* .

Grimmer, J. & King, G. (2010), 'General purpose computer-assisted clustering and conceptualization', *PNAS* .

Hopkins, D. & King, G. (2010), 'A method of automated nonparametric content analysis for social science', *American Journal of Political Science* .

Hopkins, D., King, G., Knowles, M. & Melendez, S. (2013), *Pacchetto ReadMe: software for automated content analysis.*
URL: *http://gking.harvard.edu/readme*

King, G. & Lu, Y. (2008), 'Verbal autopsy methods with multiple causes of death', *Statistical Science* pp. 78–91.

Lowe, W. (2008), 'Understanding wordscores', *Political Analysis* .

Meloncelli, D. (2012-2013), 'Sentiment analysis in twitter', *Università degli studi di Bologna* .

Oppici, Martin, D., Morando, Basso & Futia (2014), 'Social university: Le università italiane sui social network', *Nexa Working Paper* .

Pang, B., Lee, L. & Vaithyanathan, S. (2002), 'Thumbs up? sentiment classification using machine learning techniques.', *In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)* pp. 79–86.

Porter, M. F. (1980), 'An algorithm for suffix stripping', *Program* pp. 130–137.

R, C. T. (2013), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.
URL: *http://www.R-project.org/*

Satta, M. (2014-2015), 'Nuove strategie per la sentiment analysis applicata ai social network', *Politecnico di Milano* .

Solari, D., Finos, L., Redaelli, M., Rinaldo, M., Branca, M. & Ferraccioli, F. (2013), *Pacchetto TextWiller : collection of functions for text mining, specially devoted to the italian language.*

Turney, P. D. (2002), 'Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews', *National Research Council of Canada* .

Wickham, H. (2014), *Pacchetto RSQLite: SQLite interface for R.*