# UNIVERSITA' DEGLI STUDI DI PADOVA
## Facoltà di SCIENZE STATISTICHE

CORSO DI LAUREA IN STATISTICA E TECNOLOGIE INFORMATICHE

Relazione Finale di Laurea Triennale

# Costruzione di un vocabolario italiano per la Sentiment Analysis e valutazione della performance sui tweets del Festival di Sanremo 2010

Relatore: Finos Livio
Correlatore: Solari Dario

Laureando: Cauduro Federico
Matricola: 554355 – STI

Anno Accademico: 2009/2010

# Indice :


| | Pagina |
|---|---|
| **INDICE** | **3** |
| **Introduzione** | **5** |
| **CAPITOLO I** | |
| **CONSUMATORI E INTERNET** | **7** |
| 1.1 La voce dei consumatori | 8 |
| 1.2 Microbloggin | 9 |
| **CAPITOLO II** | |
| **SENTIMENT ANLAYSIS** | **10** |
| 2.1 Lo Studio | 11 |
| **CAPITOLO III** | |
| **IL VOCABOLARIO E LO DELLA SENTIMENT ANALYSIS** | **12** |
| 3.1 Il Vocabolario | 12 |
| 3.1.1 Lo stemmer | 15 |
| 3.2 Alcune Funzioni Sviluppate e comandi utilizzati | 16 |
| 3.2.1 Pulisci Tweets.R | 16 |
| 3.2.2 Keybrands.R | 17 |
| 3.2.3 Creazione di liste per la polarità | 18 |
| 3.2.4 Utilizzo della funzione Dictionary( ) | 19 |
| 3.2.5 Estrazione degli URL | 20 |
| 3.2.6 Alcuni comandi utilizzati | 20 |

**CAPITOLO IV**

**L'applicazione** **23**

4.1 La Sentiment Analysis 24

4.2 Il confronto 29

**Conclusioni** **36**

**BIBLIOGRAFIA E SITOGRAFIA** **37**

# Introduzione

La Sentiment Analysis è la capacità di decifrare le opinioni contenute in un testo scritto o parlato e da la possibilità di estrarre, tramite i computer, queste opinioni da qualunque Social Media in modo automatico utilizzando elementi di machine learning come l'analisi latente semantica. E' un nuovo approccio che determina l'atteggiamento di un oratore o uno scrittore rispetto a qualche argomento, ed il compito fondamentale è la classificazione della polarità di un'espressione lessicale, la quale può avere polarità neutra, positiva o negativa. Con questo sistema, si è in grado di identificare automaticamente la polarità delle espressioni e quindi si possono prendere delle misure di gestione della conversazione in Rete che sono dette "social customer relationship management", un'evoluzione del "customer relationship management" utilizzato da anni per la gestione del cliente.

Il problema di questo nuovo studio è che i fattori culturali, linguistici e le sfumature nei differenti contesti rendono estremamente difficile la trasformazione di una stringa di testo scritto in un semplice sentimento pro o contro.

Nel seguente elaborato si andranno a studiare dei post pubblicati da utenti nel sito di Twitter riguardanti la partecipazione di Sanremo 2010 dei cantanti Scanu e Malika. Codesto sito internet è un servizio gratuito di social network e microblogging che fornisce agli utenti una propria home page aggiornabile tramite i messaggi (SMS, email, applicazioni che utilizzano una tecnologia API ) di una lunghezza massima di 140 caratteri, e tali aggiornamenti sono istantanei, ovvero la prima pagina viene aggiornata con l'immediato inserimento di essi.

I social network consistono in un gruppo di persone connesse tra loro da diversi legami, mentre i microbloggin che vedremo nei prossimi capitoli, è una forma di pubblicazione costante di brevi contenuti di testo, immagini, file audio che verranno inseriti in un servizio di social network, e Twitter è proprio uno di questi microblogging.

Nella prima parte della tesina seguirò a presentare il comportamento dei consumatori in internet, facendo una breve descrizione della variabile voce, che è molto importante nell'ambito del marketing, e di cosa sono i microbloggin.

In seguito vi sarà una breve parte dedicata alla spiegazione della Sentiment Analysis, che cos'è, perché viene utilizzata e come costruirla passo per passo.
Nella terza parte viene applicato il lavoro svolto ai micro post estrapolati dal sito internet sopra citato, attraverso tale analisi cercheremo di valutare le capacità previste del vocabolario costruito per arrivare a delle conclusioni che potrebbero risultare utili a capire meglio gli atteggiamenti delle persone.
I fini principali di questa ricerca sono due: il primo la creazione di un vocabolario d'italiano che sarà poi utilizzato nell'applicazione dell'analisi per studiare i post, il secondo è di valutazione della performance dello strumento ottenendo naturalmente un esito positivo nel funzionamento dell'applicazione, questa si potrà applicare successivamente nell'ambito del marketing alle decisioni aziendali per portare miglioramenti ai propri prodotti nel mercato.

# CAPITOLO I

## Consumatori e Internet

Internet ha un notevole impatto sulle strategie che il consumatore può adottare per risolvere la problematicità relativa all'acquisto, che genera un fabbisogno informativo che la rete contribuisce a soddisfare ospitando:

a. un gran numero di fonti informative relative ai prodotti e al loro utilizzo;
b. contesti virtuali;
c. strumenti e siti che assistono il consumatore nel processo di acquisto e agevolando l'analisi comparativa delle alternative.

L'aspettativa del consumatore è di poter esser supportato non solo nelle fasi di ricerca e selezione delle informazioni ma anche in quella di valutazione. Ciò si realizza attraverso agenti elettronici e siti collettivi come quelli del consumerismo e le comunità virtuali di consumatori, che si raffigurano come spazi aperti per la ricerca e lo scambio di informazioni, la condivisione di esperienze, la sperimentazione e la produzione di conoscenze portando così l'apprendimento derivante dall'esperienza d'uso dei prodotti.
Un costrutto utile per lo studio del comportamento in internet è quello di set mentale (mindset).
I consumatori-navigatori possono essere suddivisi in tre principali categorie:

- Goal Oriented (GO) : orientati all'obiettivo di acquistare un determinare prodotto ed interessati a percorrere le fasi di acquisto online, consumatori “transazionali”.
- Experiantial Consumer (EC) : consumatori dotati di un mindset orientato all'esperienza, e sono rappresentati dai collezionisti e possono essere denominati anche consumatori “relazionali” .
- Tra queste due categorie c'è un'area intermedia formata da utilizzatori abituali di internet, per i quali è del tutto normale inserire nella fase informativa la visita di un sito internet specifico o un'ampia esplorazione della rete
  [Roberto Grandinetti – Marketing, mercati prodotti e relazioni].

La collaborazione e la comunità sono importanti caratteristiche dello sviluppo del Web 2.0, e sono caratteristiche chiave dei servizi di comunicazione come i siti di social network (ad esempio, MySpace, Facebook e LinkedIn), virtual reality (ad esempio Second Life), e comunità online (ad esempio, Wikipedia, YouTube e Flickr).
Grazie a questi social network si possono esaminare gli effetti dei servizi del settore commerciale, cioè, l'impatto del rapporto che viene a crearsi tra azienda e cliente, perché essi tramite il meccanismo denominato word of mouth brading, ovvero passaparola, possono modificare le idee della gente.

## 1.1 La voce dei consumatori

Le persone fanno affidamento sulle famiglie, gli amici, e gli altri nel social network. La ricerca indica anche che le persone sembrano fidarsi di pareri apparentemente disinteressati di persone al di fuori del loro immediato social network, come ad esempio riviste online; questa forma è conosciuta come onlineWOM (OWOM) o electronicWOM (eWOM).
Più precisamente il passaparola (WOM, in marketing voce del consumatore) è il processo di trasmissione delle informazioni da persona a persona e costituisce una forma particolare di impegno nella fase successiva all'acquisto. In situazioni commerciali, il WOM può modificare gli atteggiamenti, le opinioni, o le reazioni da parte dei consumatori su aziende, prodotti o servizi con altre persone.
Il potenziale di voce dipende dal numero di clienti che non hanno scelto, o non ancora, l'opzione uscita, ovvero il cliente cessa di comprare i prodotti dell'azienda.
La voce comporta in genere dei costi per l'impresa, ma se il potenziale di voce si traduce in voce effettiva ed il management ha sviluppato un'adeguata capacità di ascolto e risposta, la voce innesca il processo di recupero dove il management può porre rimedio all'errore commesso [Roberto Grandinetti – Marketing, mercati prodotti e relazioni].
Una forma potenzialmente nuova di marketing sono gli eWOM microblogging i quali poi si possono analizzare tramite una Sentiment Analysis che vedremo nel prossimo capitolo.

## 1.2 Microbloggin

I microbloggin sono mini post che vengono distribuiti per messaggi istantanei tramite telefoni cellulari, e-mail o web, e la lunghezza standard di uno di essi è paragonabile circa alla lunghezza di un sottotitolo dei giornali.
Questi blogs utilizzano i servizi della comunicazione tramite web come Twitter, dove gli utenti possono descrivere le cose di interesse ed esprimere atteggiamenti e “sentimenti” verso i prodotti di una marca e/o di un'impresa che essi sono disposti a condividere con gli altri utenti del web.
Il Microblogging è indicato anche come micro-sharing, il micro aggiornamento, oppure Twittering (da Twitter,di gran lunga l'applicazione di microblogging più popolare).

# CAPITOLO II - Sentiment Analysis

La Sentiment Analysis ha ricevuto, solamente di recente, l'attenzione da parte della comunità accademica, e solo negli ultimi anni si è vista una piccola esplosione di pubblicazioni per esempio se ne è occupata la Teragram, divisione di SAS e leader nel settore del Natural Language Processing (NLP) con l'annuncio del lancio della SAM (Sentiment Analysis Manager ), il NY Times che spiegò come grandi aziende analizzano le fonti dei social network ed in particolare Twitter per capire cosa dicono i loro clienti, ed anche google ha creato un'applicazione chiamata google sentiment attraverso la quale si possono ricercare indagini effettuate proprio con questo approccio.

La maggior parte degli studi svolti è stato fatto a livello di documento, ad esempio distinguendo le recensioni positive da recensioni negative. Questa nuova tecnica è importante per le ricerche di marketing, in cui le imprese vogliono scoprire cosa la gente pensa dei loro prodotti, tramite il monitoraggio dei newsgroup e forum, nei quali la rilevazione di feedback dei clienti è rapida.

L'analisi viene formulata su dati estrapolati dal web, i quali sono importanti per ogni azienda/istituto che cura il controllo di qualità.

I problemi che si potrebbero riscontrare sono di creare sondaggi in maniera adeguata, con adeguate domande, per ciascun prodotto o per le funzionalità d'essi, il supporto, la distribuzione del sondaggio e per i tempi e costi elevati, ma grazie ad internet potremo considerare questi problemi obsoleti, perché tramite il World Wide Web si possono estrarre un enorme quantità di informazioni che rispondono a tali domande.

Una delle tante fonti di codeste informazioni sono i blog, dove il proprietario del blog commenta un determinato argomento o parla per esperienza personale stimolando i lettori a partecipare anch'essi alla discussione o scambio di propri pareri. Un'altra fonte sono i forum dove, diversamente dai blog, ogni persona può discutere qualsiasi argomento o chiedere un parere/consiglio, per esempio quale notebook tra mac e sony è il migliore, di altre persone.

Ci sono anche dei vantaggi in questo, in primo luogo, le persone che condividono le loro opinioni di solito sono coloro che hanno pareri più pronunciati influenzando la lettura dei “curiosi” e determinando in tal modo il così detto word-of-mouth (WOM, visto nel

capitolo 1). Il secondo vantaggio, ma non per questo il meno importante, è che le opinioni possono venire estratte in tempo reale, consentendo in questo modo alle aziende di avere un tempo di reazione più rapido ai cambiamenti del mercato e alle statistiche dettagliate per poter supervisionare l'evoluzione di esso nel tempo.

## 2.1 Lo studio

Prima di effettuare la classificazione dei "sentimenti", dobbiamo individuarli all'interno dei tweets, che in genere sono emozioni oppure giudizi o idee, dei quali tramite un'accurata valutazione si verifica se c'è la presenza di qualcosa di negativo o positivo, prestando particolare attenzione anche ai piccoli particolari del lessico.
Lo studio delle emozioni può essere svolto attraverso la verifica d'influenza delle emozioni di colui che scrive tramite la scelta delle parole scelte e/o degli elementi linguistici e/o dall'impostazione del testo.
Oltre a questi due modi per effettuare l'analisi, ce ne sono altri ma che non fanno strettamente parte di quelli sopra citati, e sono, per esempio, l'uso dell'ironia e quello di un linguaggio figurativo. Però è opportuno sottolineare che le maggiori tecniche di Sentiment Analysis non si basano su emozioni reali ma sulla valutazione, apprezzamento o giudizio perché si concentrano su recensioni fatte su film, prodotti, automobili, etc., dove le valutazioni sono prive di "sentimentalismi" proprio perché sostanzialmente l'utente che ha una conoscenza dell'oggetto preso in esame, da una propria valutazione in maniera distaccata, dando ai semplici lettori e/o a coloro che sono alla ricerca di confronti tra prodotti nella fase di pre acquisto l'opportunità di farsi una propria idea del medesimo oggetto.

# CAPITOLO III

# Il vocabolario e lo sviluppo della Sentiment Analysis

Come per ogni analisi che si rispetti, prima di poterla far "girare" per ottenere dei risultati, bisogna iniziare a costruirla dal principio, ovvero nel nostro caso dalla costruzione del vocabolario, come citato nell'introduzione, perché è/sarà il primo vocabolario italiano, per lo meno di libero accesso *(freeware),* il quale è stato strutturato utilizzando un vocabolario inglese, dal quale sono state tradotte le parole in modo tale di avere già una buona base di partenza.
Nei seguenti sotto capitoli verrà illustrato, in maniera più o meno sintetica, come poter svolgere l'analisi passo a passo, anche con le parziali costruzioni delle funzioni che si andranno poi ad utilizzare oltre al vocabolario e quelle ad esso collegate.

## 3.1 Il vocabolario

Il vocabolario è sostanzialmente un elenco di termini classificati secondo il sentimento positivo o negativo che esprimono.
In questo lavoro i dati sono strutturati in un data frame composto da sei colonne dove si ha nella prima la distinzione a seconda se la parola è *weaksubj*, ovvero quelle parole che sono debolmente soggettive, o *strongsubj*, parole che nella maggior parte dei contesti sono fortemente soggettive.
Nella seconda colonna sono presenti le parole del dizionario inglese, mentre nella terza le corrispondenti e svariate traduzioni in italiano, poi abbiamo nella quarta lo *stemmed* delle parole italiane, nella penultima colonna si ha di che tipo è la parola, cioè se è un aggettivo, nome, verbo, ed in fine nell'ultima si ha la corrispettiva polarità di ogni singola parola (Positive, Neutral, Negative e Negation, dove queste ultime sono le negazioni forti che invertono la polarità o la intensificano).
In un'analisi, molti piccoli particolari ma di cruciale importanza devono essere presi in considerazione, ad esempio bisogna prestare particolare attenzione alle negazioni, che possono essere locali, oppure a "distanza", cioè negano la preposizione o il soggetto

(es. nessuno pensa che sia buono → non piace a nessuno → polarità della frase negativa), quindi bisogna invertire la polarità, ma le negazioni servono anche come intensificazione di una negazione, inoltre l'attenzione deve focalizzarsi anche nel senso delle parole, il ruolo sintattico di una particolare parole inserita nel contesto della frase.
Le colonne denominate type, eng_word, ita_word, kind e priorpolarity sono state realizzate "a mano", ovvero scritte una ad una utilizzando il vocabolario cartaceo inglese-italiano, la colonna ita_stem è stata sviluppata tramite R, con funzioni presenti nelle librerie RWeka, Snowball e tm ed un piccolo controllo di sostituzione di alcune vocali che vengono ripetute di seguito in unica parola (esempio, bellooooo->bello)

```
tweets.stem<-function(file)
{
    file <- WordTokenizer(file)
    file <- SnowballStemmer(file, control=Weka_control(S="italian"))
    file <- paste(file,collapse=" ")
}
```

Per prima cosa si carica il dizionario

```
sentimentDB<-read.csv2("Sentiments_DB1.csv",sep = ";",na.strings = "")
```

viene richiamata la funzione tweets.stem passando solamente la colonna ita_word, specificando che sono di tipo character e che dove mancano i valori (NA) devono essere sostituiti con " ".
In fine si scrive il vocabolario aggiornato con l'aggiunta della nuova colonna ita_stem.

| | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| 1 | TYPE | ENG_WORD | ITA_WORD | ITA_STEM | KIND | PRIORPOLARITY |
| 2 | strongsubj | abandoned | abbandonati | abbandon | adj | negative |
| 3 | strongsubj | abandon | abbandono | abband | noun | negative |
| 4 | strongsubj | abandon | abbandonare | abbandon | verb | negative |
| 5 | weaksubj | abandon | lasciare | lascia | verb | negative |
| 6 | strongsubj | abandonee | abbandonatario | abbandonatar | noun | negative |
| 7 | strongsubj | abandonament | abbandono | abbando | noun | negative |
| 8 | weaksubj | abandonament | rinuncia | rinunc | noun | negative |
| 9 | weaksubj | abandonament | recesso | recess | noun | negative |
| 10 | weaksubj | abandonament | desistenza | desistenz | noun | negative |
| 11 | strongsubj | abase | degradare | degrad | verb | negative |
| 12 | weaksubj | abase | abbassare | abbass | verb | neutral |
| 13 | strongsubj | abase | umiliare | umilmen | verb | negative |
| 14 | strongsubj | abasement | mortificazione | mortif | noun | negative |
| 15 | strongsubj | abasement | umiliazione | umil | noun | negative |
| 16 | strongsubj | abasement | degradazione | degrad | noun | negative |
| 17 | weaksubj | abash | confondere | conf | verb | negative |
| 18 | weaksubj | abash | imbarazzare | imbarazz | verb | negative |

Fig.1 Il vocabolario

Il problema principale riscontrato in questa fase è stato nell'assegnare la polarità alle parole, perché alcune di esse possono essere tutte e tre le cose (positive, negative o neutral), a seconda della loro sistemazione nella frase, dall'interpretazione assegnatale che può variare da persona a persona (esempio: Luigi frega l'auto, la parola frega può avere due significati, o viene intesa come strofinare l'auto, e può essere positiva o neutrale, oppure come rubare l'auto, in questo caso di certo assume polarità negativa), un ulteriore problema sono le svariate traduzioni dall'inglese all'italiano, ma anche il fatto che se le parole straniere sono prese dall'americano e non dal puro inglese hanno ulteriori definizioni.

### 3.1.1 Lo Stemmer

Lo stemming è il processo di riduzione della forma flessa, ovvero una qualsiasi variazione morfologica delle parole realizzata per indicarne i tratti grammaticali o sintattici, di una parola alla sua forma radice, detta tema. Il tema non corrisponde necessariamente alla radice morfologica della parola: normalmente è sufficiente che le parole correlate siano mappate allo stesso tema, ad esempio: 'gatto', 'gatta', 'gattino', 'gattaccio' si vorranno mappati sul tema 'gatt'.

Il primo stemmer fu proposto nel 1968 da Lovins, poi migliorato da Martin Poter nel 1980, dove la sua codifica originale fu in BCPL, un linguaggio all'epoca popolare. Questa tecnica viene studiata nell'ambito dell'informatica da ben quarant'anni ed è molto usata perché è uno dei metodi che permette di ridurre le dimensioni dei documenti e quello che ci si aspetta è che lo stemmer riconosca, in modo più accurato possibile, la correlazione tra i termini a cui viene attribuita una semantica comune e che sfrutti questa per sostituire con il tema corrispondente tutti i termini correlati.

Gli algoritmi di stemming possono venire classificati in famiglie a secondo del campo di utilizzo, e possono essere algoritmi a forza bruta, che operano per rimozione di suffissi (suffix stripping), di lemmatizzazione, stocastici, ibridi e di matching.Nella nostra ricerca abbiamo usato un algoritmo di forza bruta, i quali utilizzano una grossa tabella di lookup per memorizzare l'insieme delle coppie delle parole flesse e le radici di costoro, la tabella viene fornita dall'uomo come nel nostro caso.

Un vantaggio è che il miglioramento dello stemmer in questo caso non richiede alcun sforzo, né particolari competenze linguistiche, perché basta integrare i nuovi casi nella tabella. Ma come i vantaggi, ci sono anche dei svantaggi ovvero, l'accuratezza dipende dalla dimensione della tabella, la sua costruzione richiede un processo dispendioso e tutt'altro che banale e deve essere costantemente aggiornata con nuovi termini. Un altro punto a sfavore è dato dai vocaboli diversi di significato ma che attraverso lo stemmer assumono radici uguali, ad esempio le parole bravo e bravata entrambe vengono “stemmate” in brav, allora c'è stato l'immediato ricorso all'uso dello stemming ad hoc il quale comporta ad avere la parola bravata scritta come bravat nella colonna ITA_WORD

e aggiungendo un'ulteriore colonna d'informazione nella quale vengono riposti tutti i modi possibili con la quale può essa terminare in modo tale che poi l'analisi “non faccia confusione” tra termini e polarità.

## 3.2 Alcune Funzioni Sviluppate e comandi utilizzati

### 3.2.1 Pulisci Tweets.R

Il problema dei post è che gli utenti li scrivono aggiungendo faccine, chiamate anche emoticons, attraverso le quali vengo espressi dei sentimenti (un esempio di codesti simboli è la classica faccina col sorriso :-) ), e queste non appartenendo al vocabolario devono essere studiate in qualche maniera, e per poter effettuare lo studio al meglio è stata strutturata la funzione pulisci tweets.
Tale funzione viene fatta prima dello *stemming*, e prende in esame tutte le righe di testo estrapolando i vari emoticons e sostituendoli con il nome EMOTEGOOD, EMOTELOVE, EMOTEBAD e via dicendo. Tali nomi non comprendono solo un emoticons ma tutti quelle che appartengono a quella categoria, ad esempio EMOTEGOOD comprende :) , :-) , 8D , :D , :-D , etc.
Per poter svolgere tali passaggi sfruttiamo la funzione gsub() di R, la quale viene utilizzata nel seguente modo:

```
file <- gsub(
"\\:\\)|\\:\\-\\)|\\:\\]|\\:\\-\\]|\\=\\)|\\=\\]|\\=\\>|\\:\\>|\\^\\^|\\^\\_\\^|\\^\\-\\^|\\^\\o\\^",
' EMOTEGOOD ', file, perl=T)
```

gsub() ha lo scopo di ricercare all'interno di un vettore di caratteri una stringa di caratteri avente espressione regolare. Le espressioni regolari vengono utilizzate da utilities programs, editor di testo e da tutti i programmi che sfruttano la manipolazione di test, un esempio ti tali funzioni sono la regex o regexp.

In presenza di metacaratteri  (caratteri speciali), questa tipologia di funzioni fanno si che essi vengano interpretati come caratteri normali semplicemente facendoli precedere da un carattere predefinito (nella maggior parte dei casi è “\”, per esempio un segno di punteggiatura viene tradotto con qualche sequenza di caratteri particolari, ma se preceduto da “\” rappresenta se stesso).
La funzione pulisci tweets, come accennato in precedenza, oltre a sostituire le emoticons, sempre tramite la funzione gsub(), viene utilizzata anche per la ricerca dei caratteri speciali, della punteggiatura, ed in questo studio si rimuovono gli @ , gli url sostituendoli con “ “, ma anche i caratteri speciali con la sostituzione di essi in caratteri ASCII.

Esempio

```
# Carattere speciale, codice per è
file <- gsub("<U+00E8>",'è', file, fixed=T)

# Punteggiatura, in questo caso si studia il !
file <- gsub("\\!{1}",' ! ', file, perl=T)
```

### 3.2.2 Keybrands.R

Un'altra funzione utilizzata simile a quella vista in precedenza è keybrands . Anch'essa utilizza gsub(), ed è stata creata per sostituire delle parole ad esempio i nomi dei cantanti presenti nei tweets (Valerio Scanu, Malika Ayane , etc.), la decisione di fare due funzioni simili tra loro ma separatamente è stata presa per il fatto che la prima, pulisci tweets, è una funzione riutilizzabile anche per altre Sentiment Anlysis, mentre quest'ultima è inerente solamente con la nostra ricerca del Festival di Saremo 2010.

Esempi di ricerca e sostituzione

```
# viene estratta la stringa valerio scanu e sostituita con scanuartist

file <- gsub('valerio scanu',' scanuartist ', file)

# la striga scanu viene cambiata anch'essa in scanuartist

file <- gsub('scanu ',' scanuartist ', file)

# il titolo della canzone del cantante, una volta evidenziata nei post da gsub(), viene
# trascritta come scanusong

file <- gsub('per tutte le volte che',' scanusong ', file)

# due sostituzioni del nome del compositore della canzone che ha interpretato Valerio
# Scanu al Festival

file <- gsub('pierdavide carone',' scanuauthor ', file)

file <- gsub('carone ',' scanuauthor ', file).
```

### 3.2.3 Creazione di liste per la polarità

Questi codici sono stati creati per il recupero delle informazioni riguardanti le polarità dei vocaboli presenti nel dataset del vocabolario. Tramite i seguenti comandi riportati qui sotto,

```
write(sentimentDB[sentimentDB$PRIORPOLARITY=="positive","ITA_STEM"],file="
        POSITIVE.txt")
```

```
write(sentimentDB[sentimentDB$PRIORPOLARITY=="negative","ITA_STEM"],file=
        "NEGATIVE.txt")

write(sentimentDB[sentimentDB$PRIORPOLARITY=="negation","ITA_STEM"],file=
        "NEGATIONS.txt")

write(sentimentDB[sentimentDB$PRIORPOLARITY=="neutral","ITA_STEM"],file="
        NEUTRAL.txt")
```

non abbiamo fatto altro che crearci quattro file, uno per i vocaboli con polarità positiva, uno per quelli negativi, uno per i neutrali ed uno per le negazioni, richiedendo al software R, di estrapolare dalla lista delle ITA_STEM, ad esempio per la prima istruzione, tutte le parole aventi nella colonna denominata PRIORPOLARITY,
con la polarità *positive*. Naturalmente, per la creazione, prima dei comandi sopra riportati,
abbiamo caricato il SentimentDB, nome assegnatogli al nostro database definitivo con gli stemming inclusi.
Queste liste saranno poi utilizzate da ulteriori comandi per verificare se le parole che compongono i tweets, dopo che sono stati fatti elaborare dalla funzione di stemming, a che polarità appartengo per poi assegnare un giudizio finale al post preso in esame, ovvero se una volta "studiato" è una frase positiva, negativa o neutrale.

### 3.2.4 Utilizzo della funzione Dictionary()

Ulteriore semplice istruzione, che verrà fatta girare prima dello svolgimento della Sentiment Analysis, è la funzione Dictionary().
Con codesti comandi vengono caricate le quattro liste create in precedenza (si veda paragrafo 3.2.3) tramite il comando

```
POSITIVESENTIMENT <- readLines("POSITIVE.txt", encoding="UTF-8")
```

e salvate in variabili che saranno utilizzate dal comando successivo, il quale tramite la funzione Dictionary() presente nella libreria tm costruisce, come si può intuire dal nome, un dizionario da un vettore di caratteri o da un documento in formato di matrice.

```
POSITIVEDICT <- Dictionary(POSITIVESENTIMENT)
```

### 3.2.5 Estrazione degli Url

E' stata creata anche una funzione che ha il compito di estrarre gli indirizzi internet presenti alla fine dei post, tramite l'utilizzo della funzione gregexpr(), che ricerca una stringa di caratteri avente come inizio le lettere http per poi salvare da un'altra parte in caso si volesse provare a verificare con un ulteriore studio, se il numero di commenti positivi o negativi derivino in maggior quantità da un ben preciso sito oppure no.

### 3.2.6 Alcuni comandi utilizzati

Sviluppata l'applicazione, da un primo risultato della verifica delle polarità, è stato verificato che molti vocaboli presenti nei trecento tweets, estratti casualmente in precedenza, non erano stati ancora inseriti nel vocabolario, allora per poter verificare al meglio la funzionalità di questo strumento si è riscontrata la necessità di effettuare un ulteriore verifica più approfondita di tutte le parole che costituiscono i post,
ma che non erano presenti nel dizionario ottenendo tramite R un output contenente tali parole con sotto riportate le frequenze con le quali esse sono presenti in totale in quei tweets.
Il calcolo delle frequenze si è ottenuto tramite i seguenti comandi, dividendo, come primo passaggio, le stringhe in singole parole

```
file <- WordTokenizer(textStem)
```

Verifica delle parole, presenti nei tweets, mancanti nel vocabolario restituendo le parole mancanti con la frequenza di ripetizione

```
tab <- table(file)
pos <- sapply(names(tab),function(x) length(which(POSITIVESENTIMENT==x))==0)
neg <- sapply(names(tab),function(x)
length(which(NEGATIVESENTIMENT==x))==0)
neu <- sapply(names(tab),function(x) length(which(NEUTRALSENTIMENT==x))==0)
negation <- sapply(names(tab),function(x)
length(which(NEGATIONSENTIMENT==x))==0)
esterni=which((pos*neg*neu*negation)==1)
sort(tab[names(esterni)])
```

Visualizzando l'output seguente,

```
       tra                 are                  as                 dir                  ed
         4                   5                   5                   5                   5
emoteshock                 gue                 hey                 iya                 lag
         5                   5                   5                   5                   5
   musical                  my               polit               quest            sherawat
         5                   5                   5                   5                   5
      voic                   x                cant                 chi     cristicchiartist
         5                   5                   6                   6                   6
        de                 nia-                   s                vint                  al
         6                   6                   6                   6                   7
     bliss                   c                 com                 era           favorited
         7                   7                   7                   7                   7
      just                  su                terz               valer                vide
         7                   7                   7                   7                   7
       you                   2                  io                  it           malikasong
         7                   8                   8                   8                   8
     parol          scanudancer                with                  --                  le
         8                   8                   8                   9                   9
         -            emotelov       filibertoartist                  me                  si
        10                  10                  10                  10                  10
       son          noemiartist                  da                  mi                  on
        10                  11                  12                  12                  12
pupoartist                 una                2010                  to            ytwide
        12                  12                  13                  13                  13
 emotegood                  is                   o        obamaquestion                  of
        14                  14                  14                  14                  14
       del                   1                 and                  un        mengoniartist
        15                  15                  18                  18                  20
        lo                 the                dell           scanusong                   <
        21                  21                  22                  24                  25
```

Fig.2 Output delle parole mancanti nel vocabolario ma presenti nei tweets

Altri comandi d'importanza rilevante, che sono stati sviluppati e utilizzati poi nell'analisi del confronto, sono i comandi per effettuare l'inversione di polarità dove è presente una negazione forte, ad esempio si ci fosse una frase del tipo Malika è bella sicuramente la polarità sarà positiva ma nel caso in cui la frase si presentasse nel seguente modo, Malika non è bella,  c'è la parola bella con polarità positiva, come prima, ma davanti è presente la parola non, che ha polarità negativa forte, quindi al termine dello studio la frase avrà una polarità negativa. Questo vale anche nel caso contrario, e si effettua con i seguenti comandi.

```
id.inverti=which(sentiment$NEGA>1)
tempPOS <- contastemsPOS
tempNEG <- contastemsNEG
contastemsPOS[id.inverti] <- tempNEG[id.inverti]
contastemsNEG[id.inverti] <- tempPOS[id.inverti]
```

Nel nostro caso c'è da fare un'importante osservazione, ovvero quando viene richiamata l'inversione di polarità, non è ancora presente il controllo per verificare a quale parola la negazione forte è affiancata quindi tutti i vocaboli positivi e negativi all'interno del post preso in esame vengono invertiti di polarità.

# CAPITOLO IV
# L'applicazione

Dopo aver effettuato il recupero di 3805 post pubblicati nel sito Twitter, dove i soggetti sono Scanu e Malika durante il periodo della trasmissione di Sanremo 2010, sempre tramite R, sono stati salvati in file .txt diversi a seconda della data nella quale erano stati pubblicati.
Dall'estrazione dei post, è stata fatta un ulteriore selezione di 300 tweets casuali, che sono stati consegnati a due studenti delle facoltà ai quali è stato a loro assegnato il compito di dare una personale valutazione, per poi, una volta costruito il modello sui 300 post, poter applicare l'analisi anche a tutti gli altri tweets rimanenti. Le opzioni di valutazione messe a loro disposizione sono cinque, e sono: positive, negative, neutral, both e foreign, quest'ultima perché sono stati riscontrati post riguardanti l'argomento trattato ma scritti in altre lingue che non si potrebbero studiare e ci si aspetta che la Sentiment Analysis li classifichi come neutrali.
Per l'estrazione casuale dei tweets sono utilizzati i seguenti comandi,

```
selez=sample(length(textOrig),300)
write.csv2(textOrig[selez],file="TweetsCasuali2.csv")
```

dove il primo effettua la selezione casuale dei tweets salvandoli nella variabile selez, mentre la seconda istruzione è una semplice scrittura di quei post in un file chiamato TweetsCasuali2.csv, questo per poterli avere sotto mano in modo tale da operare con gli stessi in ogni momento che si accede ad R perché altrimenti si avrebbero ogni volta tweets diversi dalla volta precedente e diventerebbe difficile se non impossibile studiarli.
Fatte codeste operazioni si passa alla costruzione di funzioni per la pulizia del testo, all'estrazione degli url dai post ed alla creazione di file contenti le liste delle parole italiane stemmed per poi procedere con la Sentiment Analysis.

## 4.1 La Sentiment Analysis

Caricati i tweets si procede con la verifica della presenza degli Url nei tweets stessi per poi inserirli nell'esito finale dell'analisi, in modo tale da verificare se la presenza di tali vada ad incidere nella polarità della frase. Fatto questo accorgimento, deve essere richiamata la funzione "pulisci tweets" (vista nel capitolo 3.2.1), per procedere con la pulizia standard del testo, ed immediatamente dopo viene effettuata una seconda pulizia delle parole più ripetute che sono inerenti solamente con la presente analisi (anche questa vista in precedenza in dettaglio) e della sostituzione di vocali ripetute più volte in una singola parola con una sola (esempio, bellooooooo → bello).
Completati i primi passaggi, si prosegue con una delle parti più importanti di questa analisi che senza codesta diventerebbe ardua la classificazione, ovvero stiamo parlando dello stemming dei microbloggin per poter poi, in un secondo momento, confrontare le parole che li compongono con la colonna delle ITA_STEM (si ricorda che sono le radici delle parole italiane presenti nel vocabolario) per far si che vengano assegnate le polarità.
Successivamente con le istruzioni Corpus e tm_map, entrambi appartenenti alla libreria tm (una di quelle utilizzate), sono stati eseguiti tre semplici ma efficaci passaggi con l'utilizzo di altrettante opzioni che sono:

- tolower, che trasforma le lettere maiuscole in minuscole o viceversa;
- removePunctuation, che rimuove la punteggiatura;
- removeNumbers, per effettuare la rimozione dei numeri.

Ora serve tutto il lavoro svolto in precedenza, da quello fatto sul dizionario a quello eseguito sui post. Si inizia col caricare le quattro liste, ognuna delle quali contiene lo stemming delle parole italiane a seconda della polarità assegnatagli (positive, negative, neutral e negation), si procede con il confronto di esse con le parole presenti nei messaggi selezionati, e si continua col recupero dei svariati tipi di emoticons, questi ultimi passaggi vengono resi possibili tramite l'utilizzo della funzione denominata DocumentTermMatrix, alla quale vengono passati il corpus dei tweets e le liste sopra citate, permettendo così la creazione di un documento sotto forma di una matrice di dati,

in modo tale che vengano poi inseriti al suo interno in colonna, uno per riga, i conteggi dei vocaboli classificati dall'analisi, per esempio i *positive*, presenti in una singola riga di testo per ogni tweets. Create queste matrici dati, una per polarità e per tipologia di emoticons, si passa alla creazione del file finale, nel quale verranno inserite codeste matrici, la presenza o meno (TRUE/FALSE) degli url e i testi originali, ovvero che non hanno subito nessuna modifica di stemmig e quant'altro, per poter così verificare con comodità gli esiti, ed infine, una volta ottenuti i giudizi delle persone, è stato possibile completare il file, inserendo anche queste liste, che saranno poi utilizzate nell'analisi successiva per verificare se l'analisi creata ha prodotto risultati più o meno vicini alle decisioni prese da "menti umane".

| B | C | D | E | F | G | H | I | J |
|---|---|---|---|---|---|---|---|---|
| votoMichel | votoFede | stemsNEU | stemsNEGA | stemsPOS | stemsNEG | emotePOS | emoteNEG | emoteJOKE |
| neutral | neutral | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| neutral | neutral | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| negative | negative | 4 | 0 | 0 | 0 | 0 | 0 | 0 |
| foreign | foreign | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| positive | positive | 6 | 1 | 1 | 0 | 0 | 0 | 0 |
| neutral | neutral | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| negative | negative | 3 | 0 | 3 | 2 | 0 | 0 | 0 |
| positive | positive | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| neutral | neutral | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| neutral | neutral | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| positive | positive | 0 | 0 | 1 | 3 | 0 | 0 | 0 |
| negative | negative | 1 | 0 | 1 | 1 | 0 | 0 | 0 |
| positive | positive | 2 | 0 | 3 | 1 | 0 | 0 | 0 |
| positive | positive | 3 | 0 | 1 | 0 | 0 | 0 | 0 |
| neutral | neutral | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| negative | negative | 1 | 1 | 0 | 0 | 0 | 0 | 0 |
| negative | negative | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| neutral | neutral | 0 | 0 | 1 | 1 | 0 | 0 | 0 |
| positive | positive | 1 | 0 | 5 | 3 | 0 | 0 | 0 |
| neutral | neutral | 1 | 0 | 3 | 0 | 0 | 0 | 0 |
| neutral | neutral | 2 | 0 | 0 | 0 | 0 | 0 | 0 |
| negative | negative | 0 | 0 | 0 | 2 | 0 | 0 | 0 |
| foreign | foreign | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| foreign | foreign | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| negative | negative | 1 | 1 | 2 | 0 | 0 | 0 | 0 |
| foreign | foreign | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| positive | positive | 2 | 0 | 3 | 1 | 0 | 0 | 0 |
| neutral | neutral | 0 | 0 | 1 | 1 | 0 | 0 | 0 |
| foreign | foreign | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

Fig.3 Risultato della Sentiment Analysis

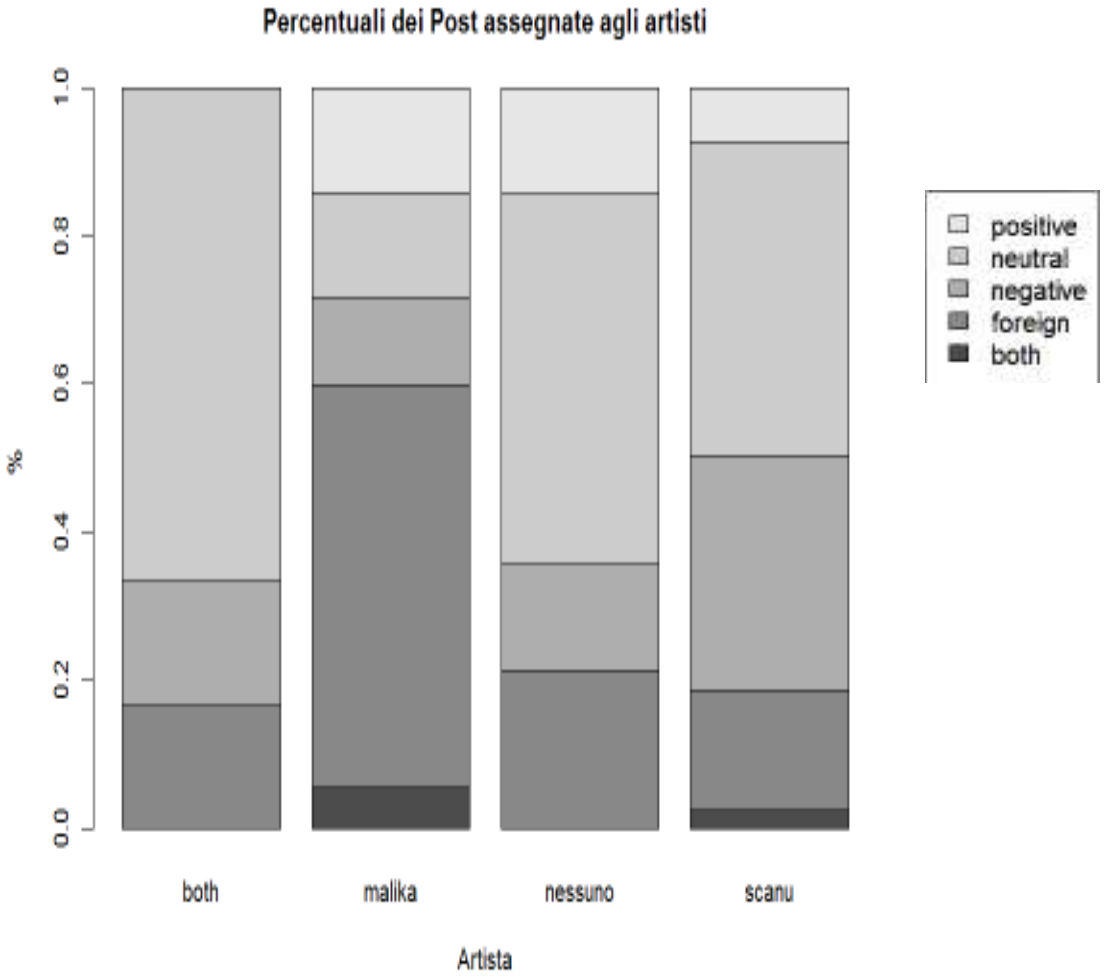
| K | L | M | N | O | P | Q | R | S |
|---|---|---|---|---|---|---|---|---|
| esiteUrl | origine | | | | | | | |
| TRUE | Vincitore Festival di Sanremo: Valerio Scanu http://bit.ly/cyW6sN | | | | | | | |
| TRUE | INDIANS @sh JUST NEWS Valerio Scanu e Alessandra Amoroso - Per tutte le volte che - | | | | | | | |
| FALSE | "RT @InWuv: #sanremo Vedo Malika e c"ho il pensiero che fa sesso con Cremonini! | | | | | | | |
| FALSE | RT @khloekardashian: House party!!! Rob, Malika, Khadijah, Jamie, Bobby, Lamar and of | | | | | | | |
| FALSE | io non ho guardato quasi per niente sanremo, e mai l'ho fatto. ho visto qualche esibizione s | | | | | | | |
| TRUE | Per tutte le volte che, Valerio Scanu a Sanremo. Video http://bit.ly/dzjz7R | | | | | | | |
| FALSE | @_nicksprincess il problema è che nn mi cambia se vince scanu o malika o noemi o pupo, | | | | | | | |
| FALSE | stendingovescion preventiva per Malika, checché ne dica dm (ms) | | | | | | | |
| FALSE | @KhloeKardashian yes she looks fab but I thought it was malika lmao. | | | | | | | |
| TRUE | • Musica - [Testo] Valerio Scanu – Per tutte le volte che http://ff.im/-guAZF | | | | | | | |
| FALSE | Vergognaaaaaaaaaaa! Ma come si fa a tenere Pupo e Emanuele Filiberto e cacciare Malik | | | | | | | |
| FALSE | @Vultrf_Tv Purtroppo devo riconoscere che fa dormire #malika #sanremo | | | | | | | |
| FALSE | #sanremo bazza, son passato davanti alla tv proprio quando c'era Malika Ayane..good!! (po | | | | | | | |
| FALSE | Vai Malika Fammi Sentire Una Bella Canzone! | | | | | | | |
| TRUE | Sanremo: Valerio Scanu Vince Il 60esimo Festival Di Sanremo http://musica.gfpoint.com/1 | | | | | | | |
| FALSE | #Sanremo Malika, non mi piacciono i tuoi capelli. | | | | | | | |
| FALSE | ha vinto Sanremo Valerio Scanu. CHI????? | | | | | | | |
| FALSE | "intanto nel backstage bosè fa la manomorta a valerio scanu, eppoi dice "scusa, m'è s | | | | | | | |
| FALSE | ... e poi finalmente spazio alle canzoni! Le mie preferite? Malika e Marco vincitore di x fact | | | | | | | |
| TRUE | Valerio Scanu vince Sanremo 2010: ascolta gratis la canzone vincitrice: Valerio Scanu, cla | | | | | | | |
| TRUE | UN GRANDE: Elio E Le Storie Tese - Per Tutte le Volte Che (Valerio Scanu Sanremo 2010) | | | | | | | |
| FALSE | RT @MarcoMm urla a gran voce: VALERIO! SCANU! VAFFANCULO! #SanRemo2010 | | | | | | | |
| FALSE | @casblake nah Malika. Her and o are bestfriends. | | | | | | | |
| FALSE | Eh dea mau lo apasih dia ga prnh jht sm lo tp lo ko kayak gitu ke malika, Cewe murahan lo | | | | | | | |
| FALSE | cmq prima si sabato notte non avevo idea di chi fosse Valerio Scanu e vivevo benissimo, ar | | | | | | | |
| TRUE | I favorited a YouTube video -- Sanremo 2010 Valerio Scanu - per tutte le volte che HQ http: | | | | | | | |
| FALSE | che bella che <U+00E8> Malika *_____* finalmente una bella voce | | | | | | | |
| TRUE | VALERIO SCANU TORNA TRA I FAVORITI http://goo.gl/fb/rmK6 #scommesse | | | | | | | |
| FALSE | E domingo foi dia de compras: Uma Melissa Malika II, Um tênis e Um vestido... ai meu bols | | | | | | | |

Per l'analisi descrittiva di questi 300 tweets, tramite alcuni controlli di ricerca di stringhe e un ciclo, è stato creato un vettore denominato artista, contenente le opzioni “scanu” e “malika” nel caso in cui, per la prima, nel tweets preso in esame compaiono le parole scanuartist, scanusong, scanu o Scanu, e per la seconda sono state ricercate le stringhe malikaartist, malikasong, malika o Malika, ma nel caso in cui alcuni post non dovessero contenere nemmeno una delle stringhe ricercate, il vettore acquisisce il valore uguale a “nessuno” e se si dovesse ritrovare una coppia di stringhe che interessano entrambi gli artisti, al vettore viene assegnato il valore “both”. La necessità di creazione di codesto vettore è stata quella di poter osservare se ci sono più osservazioni positive per Malika Ayane o per Valerio Scanu, e per tutto il resto delle votazioni. Questo studio è possibile effettuarlo creandosi la tabella sotto riportata incrociando la colonna delle votazioni con quella dell'artista  attraverso il comando table.

| | both | malika | nessuno | scanu | Totale |
|---|---|---|---|---|---|
| both | 0 | 7 | 0 | 3 | 10 |
| foreign | 1 | 69 | 3 | 20 | 93 |
| negative | 1 | 15 | 2 | 39 | 57 |
| neutral | 4 | 18 | 7 | 52 | 81 |
| positive | 0 | 18 | 2 | 9 | 29 |
| totale | 6 | 127 | 14 | 123 | |

Mentre col comando prop.table è stata realizzata la tabella contenente le percentuali di assegnazione di un artista ad una votazione, e per esempio si può notare che il 31% delle volte che si è in presenza di un tweets riguardante Scanu, esso ha polarità negativa a differenza di Malika che la stessa polarità ha una probabilità del 11%.

| | both | malika | nessuno | scanu | Totale |
|---|---|---|---|---|---|
| both | 0.00000000 | 0.05511811 | 0.00000000 | 0.02439024 | 0.07950835 |
| foreign | 0.16666667 | 0.54330709 | 0.21428571 | 0.16260163 | 1.086861 |
| negative | 0.16666667 | 0.11811024 | 0.14285714 | 0.31707317 | 0.7447072 |
| neutral | 0.66666667 | 0.14173228 | 0.50000000 | 0.42276423 | 1.731163 |
| positive | 0.00000000 | 0.14173228 | 0.14285714 | 0.07317073 | 0.3577601 |



## 4.2 Il confronto

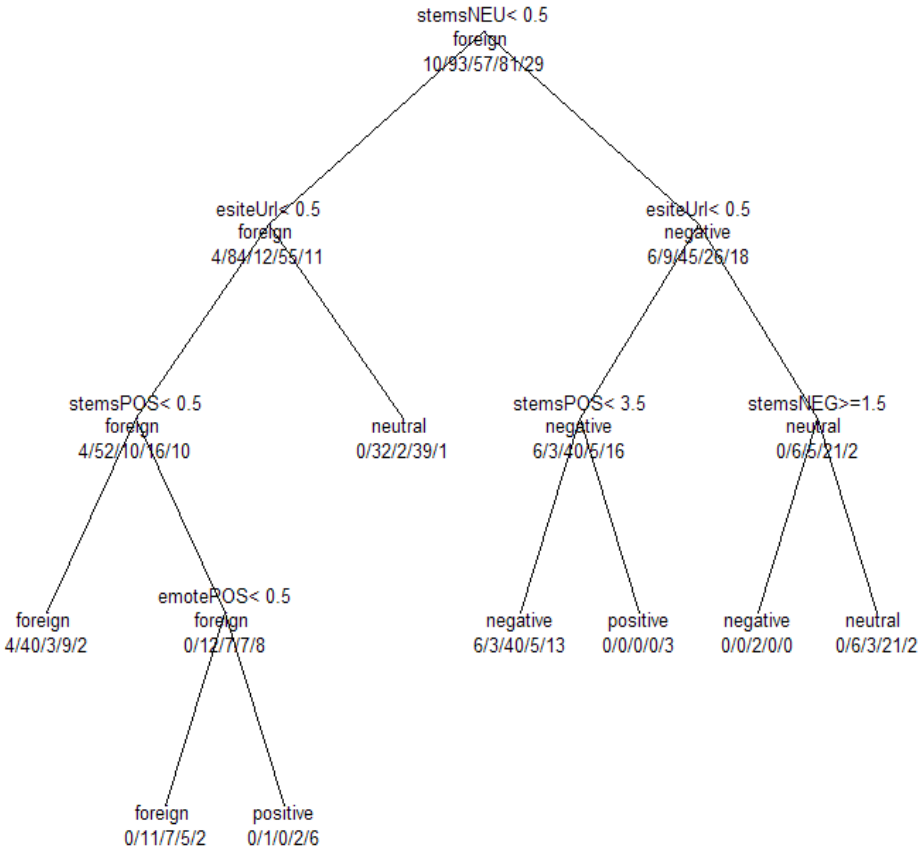
Prima di iniziare il test sulla bontà dei risultati ottenuti dalla Sentiment Analysis sono state effettuate delle modifiche al data set ottenuto precedentemente. Tali modifiche interessano le votazione effettuate dalle persone, ovvero è stato eseguito un controllo di concordanza nel giudizio e nel caso in cui i voti dei candidati fossero diversi tra loro, la riga del data set è stata rimossa da esso. Questa decisione è stata presa per il semplice motivo che lo scopo è verificare se la Sentiment Analysis arriva a calcolare un giudizio il più possibile vicino al nostro e non quale votazione, di quelle presenti, è meglio spiegata da essa. Eseguiti questi controlli prima di poter iniziare l'analisi finale è stata eseguita l'inversione di polarità delle parole, che avviene in presenza di una negazione forte, trasformando così un vocabolo positivo in negativo e viceversa (si veda paragrafo 3.2.6).
Caricata la libreria rpart, viene costruito un albero di classificazione, ove le lunghezze dei rami (verticali) sono proporzionali al guadagno corrispondente al nodo, specificando le variabile da inserire come esplicative usando il modello statistico.
Nel modello statistico scelto la variabile risposta è la votazione assegnata dall'esterno, mentre le variabili esplicative sono rappresentate dai conteggi delle parole e delle emoticons, ma anche la presenza o meno degli urls.

```
voto~esiteUrl+stemsPOS+stemsNEU+stemsNEG+emotePOS+emoteNEG+
        +emoteJOKE
```

Per spiegare una disposizione dei dati di un grafo ad albero, utilizziamo un oggetto della classe rpart che riporta oltre alle informazione sui nodi anche le varie suddivisioni effettuate, descrivendone le osservazioni campionarie del nodo. Le informazioni che vengono riportate sono: l'ultima condizione presa in esame, il numero di osservazioni nel nodo e la perdita di tali dovute alla mal classificazione, la classe prevalente ed infine la frequenza relativa se le probabilità a priori sono relative al campione. Per ottenere lo

scheletro di questa tipologia di grafo si utilizza il comando plot impostando l'attributo branch uguale a 0 e per inserire le varie condizioni che formano spiegano i rami si è fatto ricorso al comando text.

```
1)  root 270 177 foreign (0.037 0.34 0.21 0.3 0.11)
  2)  stemsNEU< 0.5 166  82 foreign (0.024 0.51 0.072 0.33 0.066)
    4)  esiteUrl< 0.5 92  40 foreign (0.043 0.57 0.11 0.17 0.11)
      8)  stemsPOS< 0.5 58  18 foreign (0.069 0.69 0.052 0.16 0.034) *
      9)  stemsPOS>=0.5 34  22 foreign (0 0.35 0.21 0.21 0.24)
        18)  emotePOS< 0.5 25  14 foreign (0 0.44 0.28 0.2 0.08) *
        19)  emotePOS>=0.5 9   3 positive (0 0.11 0 0.22 0.67) *
    5)  esiteUrl>=0.5 74  35 neutral (0 0.43 0.027 0.53 0.014) *
  3)  stemsNEU>=0.5 104  59 negative (0.058 0.087 0.43 0.25 0.17)
    6)  esiteUrl< 0.5 70  30 negative (0.086 0.043 0.57 0.071 0.23)
      12)  stemsPOS< 3.5 67  27 negative (0.09 0.045 0.6 0.075 0.19) *
      13)  stemsPOS>=3.5 3   0 positive (0 0 0 0 1) *
    7)  esiteUrl>=0.5 34  13 neutral (0 0.18 0.15 0.62 0.059)
      14)  stemsNEG>=1.5 2   0 negative (0 0 1 0 0) *
      15)  stemsNEG< 1.5 32  11 neutral (0 0.19 0.094 0.66 0.062) *
```




Dopo aver ottenuto il grafo viene richiesto ad R di creare una tabella di classificazione degli alberi, la quale riporta per ciascun grado di complessità (pari al numero di foglie meno uno) due misure di errore, espresse relativamente dall'errore del nodo radice: rel error (errore apparente) e xerror (ottenuto mediante validazione incrociata).

Root node error: 177/270 = 0.65556

| | CP | nsplit | rel error | xerror | xstd |
|---|---|---|---|---|---|
| 1 | 0.203390 | 0 | 1.00000 | 1.00000 | 0.044114 |
| 2 | 0.090395 | 1 | 0.79661 | 0.79661 | 0.046371 |
| 3 | 0.039548 | 2 | 0.70621 | 0.72881 | 0.046371 |
| 4 | 0.016949 | 3 | 0.66667 | 0.71751 | 0.046336 |
| 5 | 0.014124 | 4 | 0.64972 | 0.71186 | 0.046314 |
| 6 | 0.011299 | 6 | 0.62147 | 0.70056 | 0.046263 |
| 7 | 0.010000 | 7 | 0.61017 | 0.68927 | 0.046201 |

Estrapolata la tabella, si esegue la funzione predict, utilizzata per la predizione dei risultati dei modelli, e l'esito ottenuto è stato confrontato con la variabile risposta, tramite il comando table, il quale utilizza la classificazione cross validata per creare come output una tabella di contingenza dei conteggi delle combinazioni, per poter così vedere le differenze di classificazione tra le votazioni e la Sentiment Analysis .

| | both | foreign | negative | neutral | positive |
|---|---|---|---|---|---|
| both | 0 | 4 | 6 | 0 | 0 |
| foreign | 0 | 51 | 3 | 38 | 1 |
| negative | 0 | 10 | 42 | 5 | 0 |
| neutral | 0 | 14 | 5 | 60 | 2 |
| positive | 0 | 4 | 13 | 3 | 9 |

Fig.3 Tabella di classificazione cross validata

Dalla cross validazione si può calcolare l'errore che la Sentiment Analysis costruita può fare nel caso in cui le si faccia classificare qualcosa. L'errore si ottiene effettuando l'operazione di somma della diagonale della tabella di cross validazione diviso il numero delle righe presenti nel data set per poi sottrarlo ad 1, e l'errore ottenuto è 0.4, ovvero l'analisi ha un margine del 40% di errore dovuto alla mal classificazione.
Ma se si restringe alla sotto tabella delle sole osservazioni positive e negative

| | negative | positive |
|---|---|---|
| negative | 42 | 0 |
| positive | 13 | 9 |

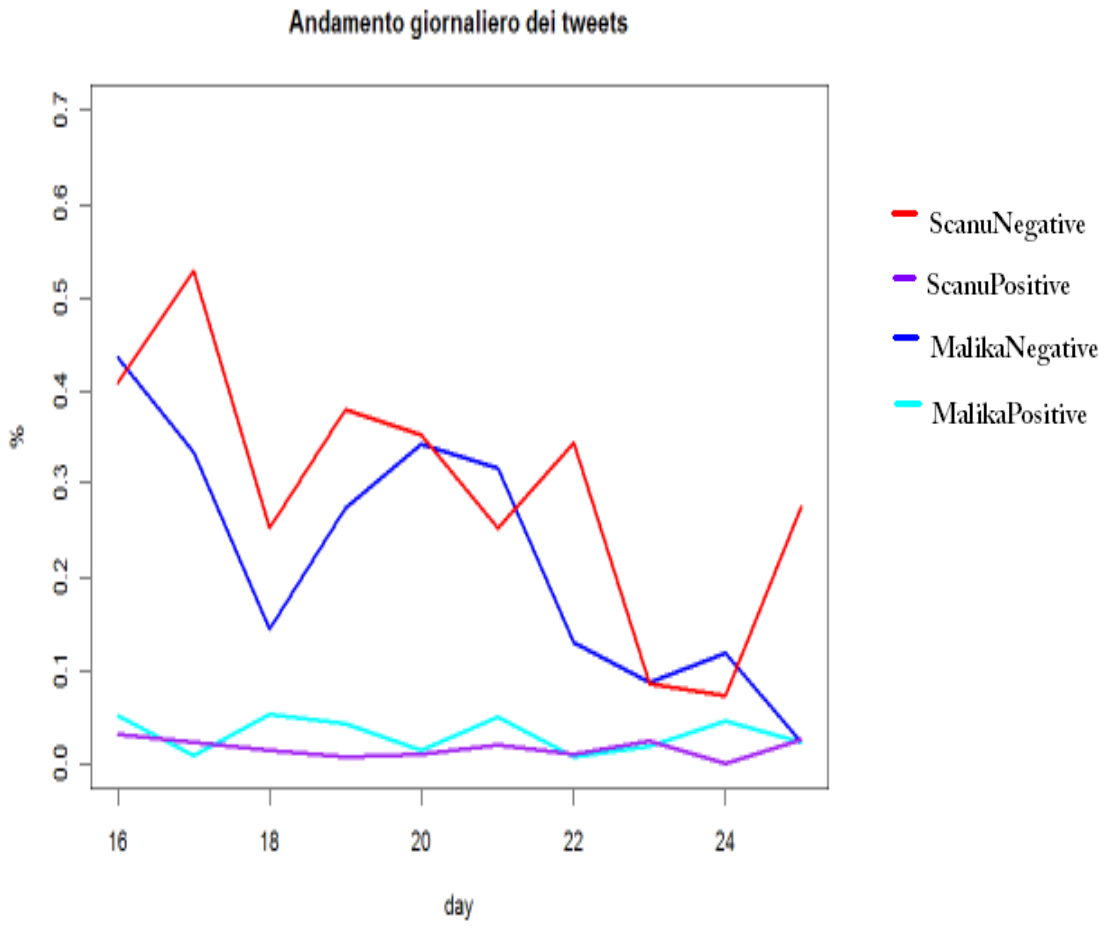
si arriva ad avere un'errore pari al 20% ricavato dalle 13 osservazioni positive classificate negative fratto la somma di queste con le osservazioni negative e con quelle positive. Questo ci porta a pensare che ci sia una sotto stima dei post positivi e una sovra stima di quelli negativi ma nell'applicazione questo è relativamente importante se ci si limita a confrontare i positivi di Malika contro i positivi di Scanu e lo stesso per i negativi.
Costruita l'analisi sui primi tweets, è stata poi applicata su tutti i 3805 che erano stati estratti, per effettuarne un'analisi descrittiva tramite una previsione delle polarità di essi per poi poterli inserire in una tabella delle probabilità avente tre dimensioni, ovvero la votazione, l'artista ed i giorni nei quali appartengono i post. Lo svolgimento di questo studio parte dal modello ottenuto in precedenza con i 300 tweets, per poi utilizzarlo nel comando predict assieme al dataset completo di tutti i post analizzati ottenendo successivamente una tabella delle votazioni previste con gli artisti ai quali vengono assegnate

| | foreign | negative | neutral | positive | Totale |
|---|---|---|---|---|---|
| both | 10 | 19 | 17 | 2 | 48 |
| malika | 709 | 408 | 471 | 52 | 1640 |
| nessuno | 140 | 55 | 105 | 4 | 304 |
| scanu | 327 | 521 | 936 | 29 | 1813 |
| Totale | 1186 | 1003 | 1529 | 87 | 3805 |

dalla quale si può notare che molti post assegnati a Malika appartengono alla classe foreign, questo perchè all'estrazione dei tweets ne sono stati prelevati anche alcuni riguardanti una certa Malika Obama per un errore di codice.

Creata la prima tabella si prosegue col comando prop.table, per poter confrontare la percentuale dei post aventi polarità positiva e/o negativa riguardanti Malika Ayane e Valerio Scanu grazie alla quale si può creare un grafico per vedere l'andamento delle dei post giornalieri a seconda della polarità prevista che i due cantanti hanno ottenuto nei giorni dello svolgimento del Festival. Questo si ottiene attraverso il comando plot e lines, dove la linea rossa rappresenta i post negativi per Scanu e quella porpora quelli positivi, mentre quella blu sono i post negativi per Malika e quella di colore ciano rappresenti i positivi.

| | 16 | 17 | 18 | 19 | 20 |
|---|---|---|---|---|---|
| Malika-negative | 0.43478261 | 0.333333333 | 0.14379085 | 0.273684211 | 0.34154930 |
| Malika-positive | 0.05072464 | 0.007751938 | 0.05228758 | 0.042105263 | 0.01408451 |
| Scanu-negative | 0.40816327 | 0.528089888 | 0.25165563 | 0.379518072 | 0.35185185 |
| Scanu-positive | 0.03061224 | 0.022471910 | 0.01324503 | 0.006024096 | 0.00925926 |
| | 21 | 22 | 23 | 24 | 25 |
| Malika-negative | 0.31578947 | 0.129251701 | 0.08620690 | 0.11818182 | 0.02272727 |
| Malika-positive | 0.04912281 | 0.006802721 | 0.01724138 | 0.04545455 | 0.02272727 |
| Scanu-negative | 0.25256410 | 0.344186047 | 0.08433735 | 0.07228916 | 0.27500000 |
| Scanu-positive | 0.01923077 | 0.009302326 | 0.02409639 | 0.00000000 | 0.02500000 |


Andamento giornaliero dei tweets
%
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
16
18
20
22
24
day
ScanuNegative
ScanuPositive
MalikaNegative
MalikaPositive

# Conclusioni

Nello studio svolto per la verifica del funzionamento della Sentiment Analysis creata, è stato costruito un modello con variabile risposta la votazione delle persone e come variabili esplicative le polarità dei tweets assegnatele dall'applicazione.

Come si è potuto notare risultato dell'errore segnala che la percentuale di possibilità di avere un'errata classificazione svolta dalla macchina è del 40%. Questo è un risultato preliminare, a prima vista può sembrare non molto soddisfacente ma bisogna tener presente che il dizionario a nostra disposizione non è ancora completo di tutti i termini presenti nella lingua italiana, che anch'essa è in continuo aumento di terminologie e di conseguenza lo stesso dizionario elettronico deve esser aggiornato in continuazione; ma oltre a questo il risultato non sbalordisce perché si è notato che a differenza dell'inglese, dal quale si è iniziato a costruire il dizionario, la nostra lingua è composta per la maggior parte di parole alle quali, prese singolarmente, viene assegnata una polarità neutrale, ma poi a seconda del contesto nel quale sono inserite, le posizioni che assumono in una frase e altre particolarità del lessico le portano ad assumere anche una polarità negativa o positiva e questo rende difficile la loro classificazione sia per un computer e sia per colui che crea/sistema/organizza il dizionario.

Un altro punto a nostro sfavore riguarda proprio i testi dei tweets estratti da Twitter, perché l'argomento trattato presenta oltre a commenti basati proprio su opinioni della gente anche commenti banali, ovvero tutti quei post fatti da siti, per esempio, dove ricordano per più volte e quindi per più post la vittoria di Scanu al festival, a differenza dei post che sono presenti in forum, blog e microbloggin dove l'argomento preso in considerazione è un costrutto di opinioni, riguardanti per esempio il confronto di oggetti, che sono molto più razionali (più tecnici che sentimentalisti).

In considerazione di tali aspetti possiamo essere soddisfati dell'analisi svolta perché nonostante questi svantaggi l'applicazione di Sentiment Analysis creata ci garantisce un 60% di possibilità di ottenere una buona classificazione e questo ci autorizza ad utilizzarla su larga scala valutando i tratti temporali, confrontando artisti o prodotti etc.

# BIBLIOGRAFIA E SITOGRAFIA

## Bibliografia

Roberto Grandinetti, 2008 - Marketing, mercati prodotti e relazioni, Carocci Editore

B.Pang and Lillian Lee, 2008 - Opinion Mining and Sentiment Analysis
http://citeseerx.ist.psu.edu/viewdoc/download?
doi=10.1.1.147.1344&rep=rep1&type=pdf

Namrata Godbol, Manjunath Srinivasaiah, Steven Skiena - Large-Scale Sentiment Analysis for News and Blogs
http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.87.7083&rep=rep1&type=pdf

Nobuhiro Kaji, Masaru Kitsuregawa – Building Lexicon for Senitment Analysis from Massive Collection of HTML Documents
http://acl.ldc.upenn.edu/D/D07/D07-1115.pdf

EriK Boiy, Pieter Hens, Koen Deschacht - Automatic Sentiment Analysis in On-line Text
https://www.law.kuleuven.be/icri/publications/892138_elpub2007.content.pdf

Bernard J. Jansen, Mimi Zang, Kate Sobel, Abdur Chowdury - Twitter Power: Tweets as Electronic Word of Mouth
http://www.mimizhang.com/paper/jansen_twitter.pdf

Wikipedia – http://it.wikipedia.org/wiki/

## Articoli

Francesco Della Rosa, 2009 - Teregram lancia Sentiment Analysis Manager al Text Analytics Summit
http://www.fullpress.it/News/Teragram-lancia-Sentiment-Analysis-Manager-al-Text-Analytics-Summit/7-32555-1.html

Alex Wright, 2009 - Mining the Web for Feelings, No Facts
http://www.nytimes.com/2009/08/24/technology/internet/24emotion.html?_r=2

Librerie R:

tm

RWeka

Snowball

rpart