G8-2001: la rivolta nel monitor. Analisi testuale dei messaggi nel newsgroup <it.eventi.g8.genova> durante gli scontri di piazza

Luca Giuliano

Dipartimento CNAPS – Università degli Studi di Roma "La Sapienza" – Italia

Abstract

Analysis of online newsgroups and chat rooms provides a rich primary source for investigation. For the most part, however, we have few tools for managing this deluge of information. These sources are very "noisy" and present us with a number of interesting technical challenges as we try to extract the useful "signal."
This paper is intended to provide an overview of a method to get a syntetic knowledge of messages in newsgroups. We analyse a political corpus obtained from an italian newsgroup on G8 during the riots in Genoa (July 20-21, 2001). A key objective for us is to develop a approach to modeling opinions about various subjects of interest. In this exemple we use S-Replace, TALTAC and SPAD-T.

Sintesi

I messaggi nei newsgroups e nelle chat rooms rappresentano una fonte di primaria importanza per la ricerca. Tuttavia, nella maggior parte dei casi, non disponiamo di strumenti adeguati per gestire questa massa ingente di informazioni. Si tratta di fonti molto "rumorose" che presentano una sfida interessante, dal punto di vista tecnico, per chi intende selezionare i "segnali" davvero rilevanti.
Questa comunicazione intende esplorare le potenzialità di un metodo in grado di fornire rapidamente una sintesi del contenuto dei messaggi. A questo scopo viene analizzato un corpus costituto dai messaggi inviati in un newsgroup italiano sul G8 durante gli scontri avvenuti a Genova il 20-21 luglio 2001. L'obiettivo è di sviluppare un approccio in grado di delineare le opinioni rispetto a vari argomenti di interesse generale. In questo esempio vengono utilizzati tre programmi: S-Replace, TALTAC e SPAD-T.

Keywords: newsgroups, textual data analysis, computer mediated communication, riots.

1. Introduzione

Vengono analizzati i messaggi inviati dal 16 al 25 luglio 2001 in un *newsgroup* nato in occasione della riunione del G8 a Genova nei giorni del 20-21 luglio. La particolare natura dell'evento, altamente mediatico e con un valore di politica internazionale di grandissimo rilievo, già nelle sue fasi preparatorie ha suscitato l'interesse del "popolo di Internet". Secondo alcuni analisti, la stessa formazione e diffusione del "movimento di Seattle" e, in generale, del "movimento anti-globalizzazione", non sarebbe stata possibile se non si fosse autoalimentata attraverso la comunicazione telematica e la fitta rete informativa nata spontaneamente tra i diversi gruppi politici e movimenti di opinione.

Il corpus da analizzare ha delle caratteristiche lessicografiche particolari e richiede un approccio metodologico sostanzialmente diverso rispetto alle procedure operative adottate solitamente nell'analisi dei testi.

Prima di tutto si tratta di un corpus ridondante, pieno di "rumore" in termini informativi e di errori di ortografia, nel quale viene utilizzato un linguaggio ibrido tra il parlato e lo scritto che rende particolarmente interessante, ma anche complessa, sia l'analisi lessicometrica che l'analisi logico-semantica.

In secondo luogo si tratta di un corpus di ampie dimensioni (9.402 Kb di testo), costituito da 8302 messaggi inviati nel *newsgroup* <it.eventi.g8-genova> dalle ore 00.01 del 16 luglio alle ore 24.00 del 25 luglio 2001.

Con un corpus di questo genere è impensabile adottare procedure non automatizzate di preparazione e analisi dei testi, sebbene ciò comporti una certa perdita di informazione e anche una notevole approssimazione nell'analisi semantica. La metodologia adottata viene proposta, quindi, attraverso un percorso operativo che dovrebbe consentire il trattamento in tempi brevi di files di testo eterogenei, rumorosi e di notevoli dimensioni.

L'obiettivo è quello di evidenziare, con un ridotto dispendio di risorse, i contenuti generali dei messaggi, i temi affrontati dagli scriventi e le loro opinioni.

2. I newsgroups

I *newsgroups* sono bacheche elettroniche, dedicate ciascuna a un diverso argomento, sulle quali gli utenti affiggono i loro messaggi o leggono i messaggi lasciati da altri. Per accedere ai *newsgroups* bisogna utilizzare un programma specifico (*newsreader*) e passare attraverso quella che può essere definita la "stanza delle bacheche", e cioè il *news server.* Ce ne sono centinaia sparsi per tutto il mondo, organizzati per gruppi linguistici. Alcuni *newsgroups* hanno un moderatore che può autorizzare o meno i messaggi da rendere pubblici.

I messaggi sono composti da una intestazione (*header*), costituita da una serie di righe di testo seguita dal "corpo" del messaggio (*body*). Ecco come appare uno dei messaggi dopo il salvataggio:

```
From: sam@infoweek.com
Date: Mon, 16 Jul 2001 00:20:06 GMT
Newsgroups: it.eventi.g8-genova
Subject: Re: Attac perquisita!!!
Organization: [Infostrada]

Filippo Vanera <vanera@yahoo.com> wrote:
> >ok..ma spiegagli anche perché !!?!?!
> Semplice: l'art. 16 della Costituzione garantisce ai cittadini il diritto
di
> muoversi liberamente sul territorio nazionale ma consente alla legge (sia
> alla legge del Parlameno, sia agli atti aventi forza di legge del
governo)
[…]

A me sembra che in questo momento ci sia un preciso intento di
"allargare" l'isolamento ben oltre la citta' di Genova, creando notevoli
disagi a tutti quelli che arrivano da altre nazioni per aderire alla
manifestazione (avendone il diritto).

Sam
++++ La coerenza e' per i VIVI.
++++ Loro sono gia' morti, solo che nessuno glielo ha detto.
```

Il linguaggio del *newsgroup* in oggetto è caratterizzato da un contesto tematico e concettuale che rimanda alla politica nazionale e internazionale, al conflitto sociale, anche molto aspro, ai più diversi ambiti della cultura (letteratura, cinema, fumetti, informazione), alla vita quotidiana e alle forme più estreme di aggressività verbale.

La condizione di stesura del testo è il risultato di una modalità asincrona di interazione, con il carattere di comunicazione pubblica da uno a molti e con uno stile in gran parte informale (sebbene vi siano messaggi che derivano da condizioni di comunicazione molto più formali, come, ad esempio, un articolo di giornale quotidiano o il comunicato di un'agenzia di stampa). In generale il carattere della comunicazione è spesso molto più polemico che informativo.

E' impossibile determinare le fonti idiomatiche della comunità degli scriventi, ma la loro collocazione socio-anagrafica è prevalentemente giovanile e i messaggi sono scritti in lingua italiana. Solo in un paio di casi sono presenti messaggi in inglese e francese. Abbiamo preferito non eliminarli dal momento che non avrebbero potuto modificare un'analisi che prevedibilmente si sarebbe limitata a prendere in considerazione solo le forme con un'alta soglia di frequenza.

Gli scriventi all'interno del *newsgroup* utilizzano un lessico in gran parte specifico della comunicazione telematica, facendo uso abbondante di *emoticon*, espressioni gergali e sigle. In generale, dal punto di vista lessicale, il linguaggio utilizzato è composito, incrociando diversi ambienti di riferimento: dalla politica alla guerra, dall'economia alla letteratura, dal linguaggio giuridico al linguaggio tecnico-scientifico.

3. La preparazione del testo e la normalizzazione

In questa prima fase l'obiettivo è individuare – il più possibile in modo automatizzato - gli elementi pertinenti all'analisi prima di passare alla normalizzazione del testo. La selezione è resa difficile dalle caratteristiche del testo, più vicino alla trascrizione di un testo orale che ad un vero e proprio testo scritto. L'interazione tra gli utenti sfugge alla convenzioni di un normale scambio di lettere; la sintassi e la grammatica spesso non vengono rispettate in funzione di una comunicazione più sintetica, più frettolosa o più polemica.

Pertanto è stata scelta una linea operativa basata sulla identificazione delle regole rintracciabili in tutti i messaggi inviati nei *newsgroups*, al fine di isolare ed eliminare le parti superflue per l'analisi. In particolare gli elementi di disturbo più frequenti sono rappresentati dagli indirizzi di posta e-mail, dagli indirizzi in Internet e dalle firme digitali.

Per automatizzare il più possibile la procedura di individuazione, eliminazione o sostituzione di questi elementi nel testo è stato utilizzato un software scritto, compilato e sviluppato sotto forma di *tool* applicativo da Alessandro Stabellini in un caso analogo per la sua tesi di laurea sull'analisi dei *newsgroups* di argomento religioso (Stabellini, 2000).

S-Replace, il programma per la "pulitura" del file originario, è stato scritto in Delphi e implementa una libreria - sviluppata e distribuita in licenza *freeware* dal russo Andrey Sorokin - che consente di sfruttare, per la ricerca del testo in un file, le potenzialità delle espressioni regolari del Perl. Il programma gira su piattaforma Windows.

Le principali sue caratteristiche sono:

- possibilità di usare le espressioni regolari del Perl per la ricerca/sostituzione del testo; - possibilità di operare su più files di testo contemporaneamente;

- possibilità di salvare i parametri di lavoro in un database editabile e richiamabile alla bisogna;

- possibilità di operare passo passo su uno o più files di testo;

- possibilità di eliminare le righe che contengono o iniziano con una determinata stringa;

- possibilità, su uno o più files contemporaneamente, di rimuovere o sostituire il testo che soddisfa determinate condizioni.

Non si riportano ovviamente tutte le sostituzioni/eliminazioni effettuate sul file oggetto di analisi. Per le firme alla fine dei messaggi, data l'impossibilità di identificare uno schema di composizione univoco (una firma è rappresentata da una porzione di testo la cui struttura è fortemente dipendente dallo scrivente), si è ritenuto opportuno eliminare solamente quelle righe di testo che iniziavano con caratteri (asterischi, tilde ecc…) ripetuti per più di 3-4 volte.

Differentemente da quanto effettuato in una ricerca simile (Beaudouin et al., 2000), si è deciso di non eliminare il *quoting*, ovvero quella parte di messaggio che l'autore decide di riportare da un altro messaggio per riferire il proprio commento. Si è ritenuto che la ripetizione di un messaggio in un discorso di carattere politico e fortemente polemico avesse l'effetto di accrescere l'impatto e l'efficacia comunicativa del messaggio stesso.

Al termine del passaggio attraverso S-Replace il file originale si è ridotto da 9.402 Kb a 6.826 Kb. La successiva normalizzazione è stata effettuata con il programma TALTAC, con il quale è stato eseguito anche il *tagging* grammaticale e la lemmatizzazione.

4. I contenuti politico-ideologici dei messaggi

Per rispettare la scelta metodologica di utilizzare la massima automatizzazione possibile della procedura senza passare attraverso una fase di controllo puntuale che avrebbe comportato inevitabilmente la lettura del testo, o almeno di una sua parte, si è deciso di procedere immediatamente alla sua elaborazione.

L'unità di contesto è rappresentata dal singolo messaggio, con una lunghezza variabile tra 1 e 250 linee di testo (128,77 occorrenze per messaggio, in media). Si tratta di una frammentazione di comodo, molto grossolana, dovuta alla impossibilità di ricorrere agli enunciati (proposizione con un senso compiuto) o alle frasi (proposizione con una rilevanza sintattica) senza una lettura puntuale dei testi.

Tab. 1 - Quadro riassuntivo con diverse soglie di frequenza

Soglia	Risposte	Occorrenze	N. forme distinte	% Forme distinte	% COP
0	8.302	1.069.035	35.766	3,35	100,00
50	8.302	872.973	2.165	0,24	81,66
100	8.302	803.935	1.178	0,15	75,20

Le parole distinte già a soglia zero sono poco numerose. Il lessico utilizzato, come si è detto, è molto vicino alla lingua parlata.

In una procedura successiva sono state eliminate le parole vuote (articoli, pronomi, preposizioni, congiunzioni), alcuni avverbi che si ritenevano poco significativi per gli scopi dell'analisi (ad esempio: sempre, invece, lì, sicuramente), e i verbi ausiliari (essere, avere). Inoltre è stato necessario fondere insieme alcune voci che erano evidentemente differenziate solo per la grafia ma non per il contenuto, benché fossero altamente rilevanti per il tema in oggetto (ad esempio: berlusconi – Berlusconi; Block – bloc - block – Bloc; ecc.). Una lemmatizzazione manuale ha permesso di includere nell'analisi alcune forme che altrimenti, avendo una frequenza inferiore a 100, sarebbero state escluse. Si tratta di forme che nella lemmatizzazione effettuata con TALTAC non sono state riconosciute dal programma o sono state classificate come ambigue.

Al termine di questa fase di selezione la matrice parole-unità di contesto (*mots-reponses*) con soglia 100, sottoposta ad analisi delle corrispondenze (SPAD-T, 1993) è risultata composta di 653 forme distinte. Nella Tab. 2 vengono riportate solo le forme "peculiari" rappresentate sul Graf. 1 selezionate in base alla rilevanza dei contributi assoluti e di un criterio di specificità del corpus rispetto al modello di riferimento rappresentato dal lessico fondamentale dei Poliformi implementato nel programma TALTAC (Bolasco, 1999).

Tab. 2 – Analisi delle corrispondenze:
coordinate e contributi assoluti delle forme peculiari.

FORME	Coordinate		Contributi ass.	
	F1	F2	F1	F2
agenti	1,08	0,51	0,4	0,1
Agnoletto	0,29	0,74	0,1	0,8
anarchici	0,84	0,54	0,4	0,2
avvocato	1,07	1,08	0,4	0,4
azienda	-1,50	0,28	0,7	0,0
Berlusconi	-0,41	0,58	0,2	0,4
black-block	1,08	0,70	2,3	1,1
camionetta	1,07	-0,77	1,2	0,7
carabiniere	0,94	-0,88	3,9	3,7
Casarini	0,64	0,67	0,2	0,3
cazzo	-0,26	-0,56	0,1	0,7
cervello	-0,24	-0,80	0,0	0,4
coglione	-0,61	-1,18	0,2	0,8
corteo	0,89	0,59	0,9	0,4
delinquente	0,66	-1,13	0,1	0,4
economia	-1,44	0,38	0,5	0,0
estintore	0,78	-1,61	0,9	3,9

fame	-1,11	0,16	0,4	0,0
fascista	-0,64	0,03	0,4	0,0
feriti	0,99	0,62	0,5	0,2
figlio	0,53	-1,70	0,2	2,6
forze_dell_ordine	0,56	0,15	0,6	0,0
foto	1,29	-0,39	1,5	0,1
G8	-0,29	0,53	0,2	0,7
Genova	0,27	0,38	0,4	0,9
Giuliani	0,94	-1,06	0,8	1,1
globalizzazione	-0,93	0,44	0,9	0,2
GSF	0,74	1,18	0,7	2,0
in_faccia	0,32	-1,24	0,0	0,5
in_mano	0,72	-0,83	0,3	0,4
insultare	-1,19	-0,35	0,4	0,0
irruzione	1,56	1,88	0,7	1,2
jeep	1,31	-1,39	0,8	0,9
lacrimogeni	1,30	0,29	0,8	0,0
lanciare	0,89	-0,51	0,4	0,1
legittima_difesa	0,67	-1,55	0,2	1,3
manifestanti	0,56	0,16	1,0	0,1
manifestazione	0,27	0,52	0,2	0,8
merda	-0,42	-1,48	0,1	1,4
messaggio	-0,79	-0,44	0,6	0,2
mondo	-0,67	0,12	1,1	0,0
morto	0,47	-0,80	0,3	1,0
movimento	-0,17	0,71	0,0	0,4
multinazionali	-1,10	0,44	0,6	0,1
omicidio	0,83	-1,16	0,2	0,4
pacifici	0,70	0,55	0,4	0,2
pianeta	-1,37	0,22	0,4	0,0
piazza	0,83	0,53	0,5	0,2
pistola	0,93	-1,39	0,6	1,4
polizia	0,75	0,61	2,7	2,0

poliziotto	0,65	0,00	0,9	0,0
portavoce	0,94	1,24	0,2	0,4
povero	0,19	-0,98	0,0	0,7
puttana	-0,14	-1,45	0,0	0,5
ragazzo	0,74	-0,76	2,0	2,3
Rifond. Comunista	0,53	1,37	0,1	0,5
rispondere	-0,55	-0,12	0,5	0,0
sangue	0,94	-0,04	0,6	0,0
sparare	0,65	-1,29	0,6	2,7
sparato	1,23	-1,68	1,3	2,7
spranga	1,00	0,23	0,4	0,0
testa	0,41	-0,81	0,2	0,8
tute-bianche	0,73	0,74	0,3	0,4
tute-nere	0,86	0,78	0,4	0,3
ucciso	0,62	-0,83	0,3	0,5
visto	0,53	0,00	0,7	0,0
vita	-0,26	-0,57	0,1	0,7

Dall'analisi emergono con evidenza quattro fattori, ma solo i primi due sono chiaramente interpretabili (Graf. 1).

Il primo fattore, sul semiasse positivo, è caratterizzato da parole come *carabiniere, polizia, black-block, ragazzo, foto, sparato, camionetta, manifestanti, corteo, poliziotto, estintore*. Il riferimento all'episodio più drammatico della manifestazione del 20 luglio è chiarissimo e ben rappresentato dalle parole indicate: la morte del giovane Carlo Giuliani che imbracciava un estintore, ucciso da un colpo di arma da fuoco esploso da un carabiniere che si trovava su una camionetta circondata dai manifestanti a Piazza Alimonda.

Il semiasse negativo è caratterizzato da parole come *mondo, globalizzazione, multinazionali, pianeta, azienda, fame,* ed esprime invece i termini essenziali del dibattito politico che è stato aperto dai partecipanti del *newsgroup* soprattutto nei giorni precedenti la manifestazione e gli scontri che l'hanno caratterizzata.

Nel complesso, il primo asse fattoriale esprime la contrapposizione tra i contenuti ideologici e politici del movimento "No Global" e del dibattito che esso produce rispetto alla narrazione epica della manifestazione, dei suoi protagonisti (il ragazzo Carlo Giuliani, i manifestanti, le Forze dell'Ordine, i carabinieri, il corteo, i Black block, le tute bianche, il Genoa Social Forum, ecc.) e dei suoi esiti più drammatici (i lacrimogeni, la jeep, la camionetta, il colpo che ha ucciso il ragazzo, il sangue, la pistola, i feriti).

Il secondo fattore è caratterizzato dal complesso delle reazioni successive alla manifestazione, sia rispetto ai soggetti politici (governativi, istituzionali, anti-istituzionali e antagonisti) che

rispetto agli episodi di violenza e alla loro gravità. Questo fattore appare come un tentativo di lettura ed interpretazione degli eventi nel momento più caldo del loro accadimento.

Il semiasse positivo è caratterizzato da forme come *Genoa Social Forum (GSF), polizia, irruzione, black-block, manifestazione, avvocato, sede, governo,*. (queste ultime due non rappresentate tra le forme peculiari). E' evidente il riferimento alla irruzione della polizia, nella notte tra il 21 e il 22 luglio, nella scuola Diaz, che era diventata la centrale operativa del movimento, e il giudizio politico sull'operato del governo.

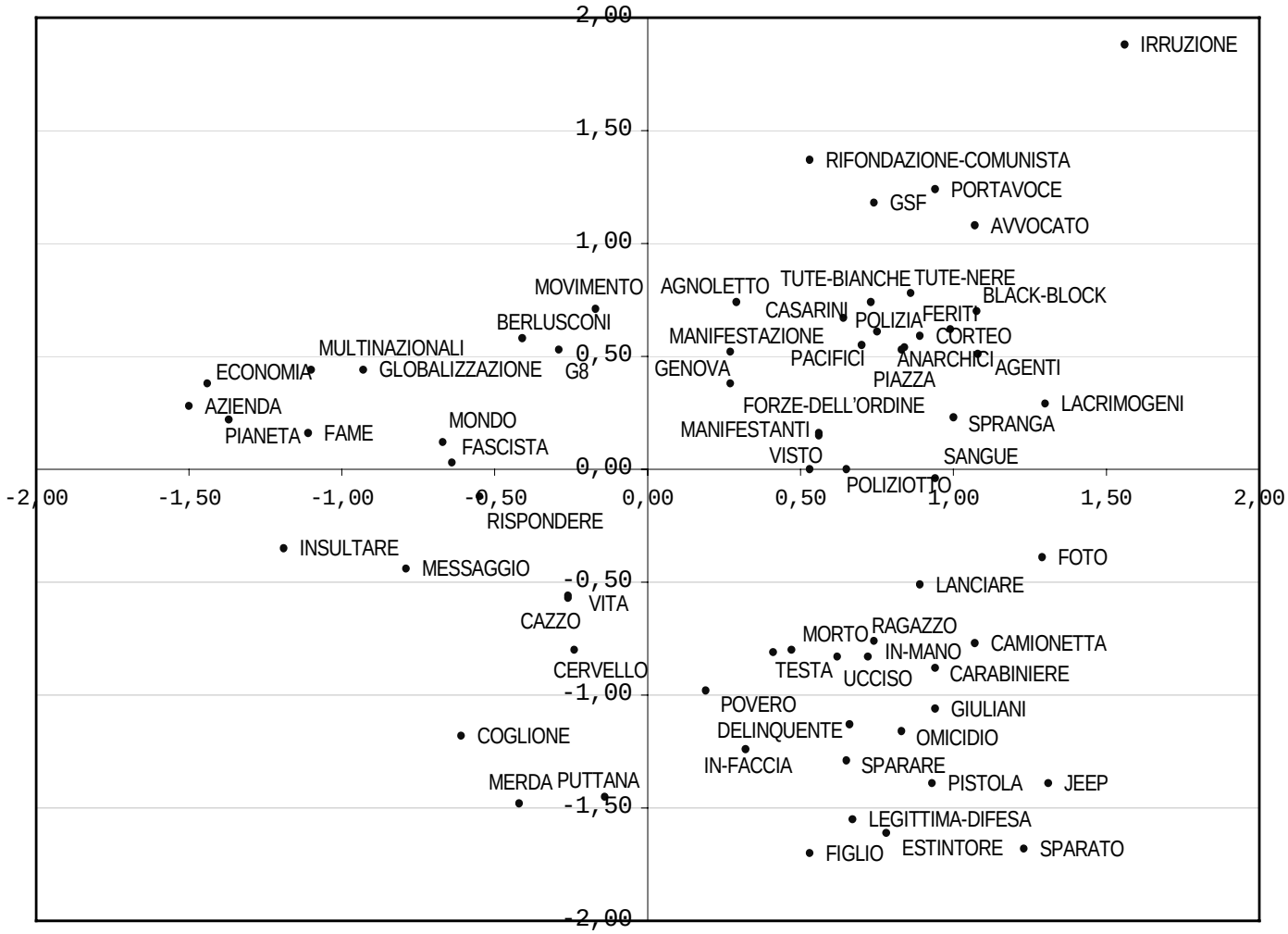
Il semiasse negativo è caratterizzato ancora una volta dalle forme che fanno riferimento alla morte di Carlo Giuliani, ma questa volta si presentano forme come *legittima difesa*, *figlio,* e soprattutto parole offensive nei confronti delle forze dell'ordine o dei manifestanti, come *merda, coglione, cazzo, imbecilli* (quest'ultima non rappresentata tra le forme peculiare).

Il piano fattoriale che emerge dall'incrocio del primo e del secondo asse evidenzia dunque tre aggregati di parole:

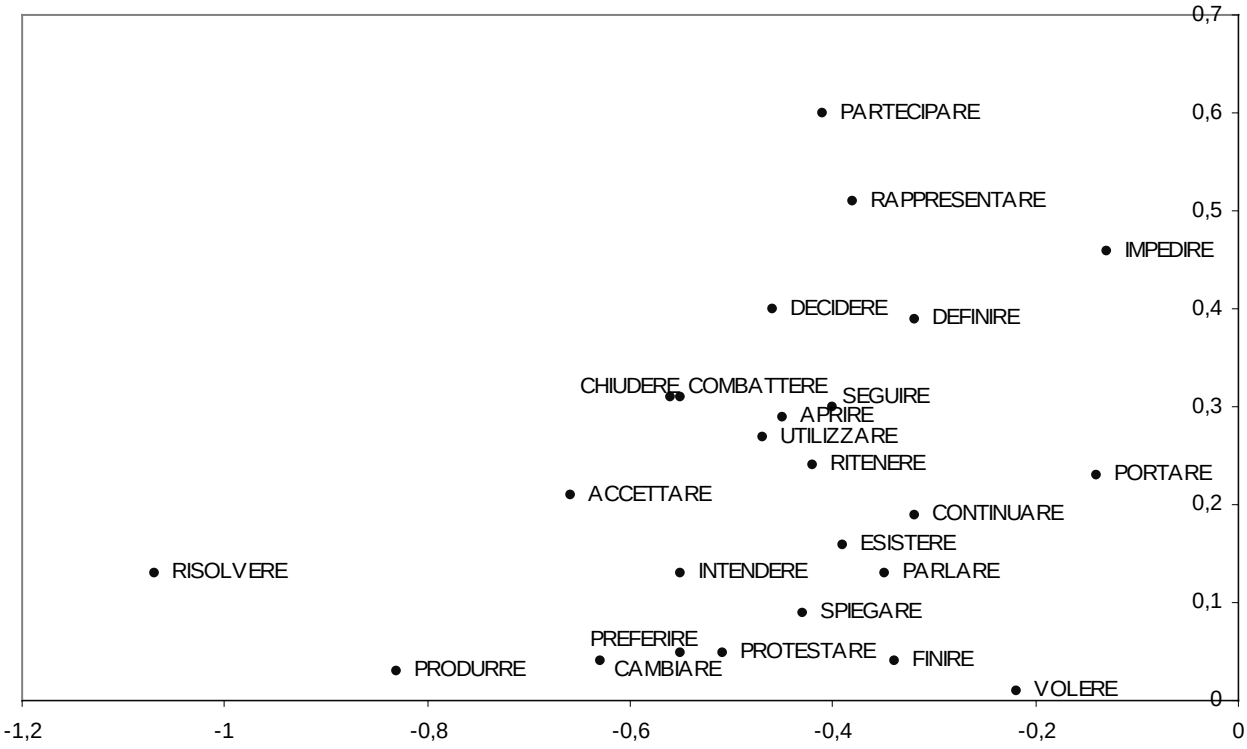
- Il primo aggregato (sul semipiano ++) si raccoglie intorno alle parole che descrivono ed interpretano l'insieme della manifestazione e le reazioni dei soggetti politici coinvolti (istituzionali ed anti-istituzionali). Il tono generale del discorso è aspro ma non provocatorio. Frequentemente è un tono di denuncia dei fatti ed esprime indignazione o consenso verso l'una o l'altra parte.

- Il secondo aggregato (sul semipiano +-) si raccoglie intorno alla descrizione degli scontri, delle devastazioni ma, soprattutto, dell'episodio che ha portato alla morte di Carlo Giuliani. Qui il dibattito diventa aggressivo e provocatorio. Le parole esprimono giudizi forti da entrambe le parti, sia di coloro che invocano una maggiore durezza da parte delle forze dell'ordine che di coloro che ne denunciano invece le violenze inutili e controproducenti.

Il terzo aggregato è rappresentato dal dibattito sui contenuti del G8, dai riferimenti alla globalizzazione, ai problemi planetari, alle ideologie, all'economia e alle multinazionali. Con una divaricazione a forbice della nuvola di punti man mano che ci si avvicina al secondo asse: da una parte si dispone il linguaggio della politica, che trova al centro i riferimenti al G8 e al presidente del consiglio Berlusconi, dall'altra il linguaggio della provocazione, dell'insulto, dello scontro verbale, che tende a collegarsi con la narrazione dello scontro di piazza e con le sue emozioni.



Graf. 1 - Proiezione delle forme peculiari sul piano fattoriale degli assi 1 e 2



Graf. 2 - Proiezione delle forme "Verbi" sul piano fattoriale dei semiassi 1(-) e 2(+)

5. Conclusione

Altre considerazioni sarebbero possibili (ma non nei limiti di questa comunicazione) esaminando in dettaglio alcune forme grammaticali che possono permettere di isolare l'intenzionalità degli scriventi e di individuare aree semantiche specifiche come: polizia, governo, Genoa Social Forum, globalizzazione. Nel Graf. 2, come esempio, si riportano le proiezioni delle forme "verbi" sul semipiano che rappresenta l'area di discussione politica intorno al G8.

In ogni caso la metodologia adottata appare come sufficientemente adeguata all'analisi e alla individuazione dei contenuti di testi di ampie dimensioni, prodotti attraverso la comunicazione telematica, sebbene con un trattamento rapido e sintetico sulle caratteristiche formali di esso. Si può supporre che la perdita di informazione sia minima rispetto all'obiettivo, ma saranno necessari altri affinamenti per poterne valutare appieno le potenzialità e gli errori.

Riferimenti bibliografici

Beaudouin, V., Fleury, S., Velkovska, J. (2000). Etudes des échanges électroniques sur internet et intranet: forums et courriers électroniques. In *JADT 2000 – 5es Journées Internationales d'Analyse Statistiques Textuelles,* Lausanne, 9-11 Mars, 2000, vol. I, pp. 17-24.

Bolasco S. (1999). *Analisi multidimensionale dei dati*, Roma, Carocci.

Bolasco, S., Baiocchi, F., Morrone, A. (2000). *TALTAC. Trattamento Automatico Lessico Testuale del Contenuto*, Roma, CISU.

Ghiglione, R., Landré, A., Bromberg, M., Molette, P. (1998). *L'analyse automatique des contenus*, Paris, Dunod.

Habert, B., Fabre, C., Issac, F. (1998). *De l'écrit au numérique. Constituer, normaliser et exploitier les corpus électroniques*, Paris, InterEditions.

Lebart, L., Morineau, A., Becue, M., Hausler, L. (1993). *SPAD-T 1.5 (DOS*), CISIA, 1993.

Stabellini, A. (2000). Il lessico dei newsgroups di argomento religioso: lo studio di quattro casi esemplari con applicazione dello SPAD-T.