Analisi Linguistica e Stilostatistica – Uno Studio Predittivo sul Campo

Rodolfo Delmonte
Dipartimento di Studi Linguistici e Culturali Comparati
Ca' Bembo – Dorsoduro 1715
Università Ca' Foscari – 30123 Venezia
Email: delmont@unive.it

Abstract

Italiano. In questo lavoro presentiamo uno studio sul campo per definire uno schema di valutazione preciso per la stilistica del testo che è stato usato per stabilire una graduatoria di diversi documenti sulla base della loro abilità di persuasione e facilità di lettura. Lo studio concerne i documenti dei programmi politici pubblicati su un forum pubblico dai candidati a Rettore dell'Università Ca' Foscari – Venezia. I documenti sono stati analizzati dal nostro sistema ed è stata creata una graduatoria sulla base di punteggi associati a undici parametri. Dopo la votazione, abbiamo creato la graduatoria e abbiamo scoperto che il sistema aveva previsto il nome del reale vincitore in anticipo. I risultati sono apparsi su un giornale locale[1].

English. *This paper presents a case study defining a precise evaluation scheme for text stylistics to be used to rank different documents in terms of persuasiveness and easyness of reading. The study concerns political program documents published on a public forum by candidates to rector of the University Ca' Foscari – Venice. The documents have been analysed by our system and a rank list has been created on the basis of scores associated to eleven parameters. After voting has taken place, we graded the different analyses and discovered that the system had predicted the name of the actual winner in advance. The result has been published on a local newspaper.*

1. Introduzione

L'analisi parte dall'idea che lo stile di un documento programmatico sia composto da elementi quantitativi a livello di parola, da elementi derivati dall'uso frequente di certe strutture sintattiche nonché da caratteristiche squisitamente semantiche e pragmatiche come l'utilizzo di parole e concetti che ispirano positività. Ho eseguito l'analisi partendo dai testi disponibili su web o ricevuti dai candidati, utilizzando una serie di parametri che ho creato per l'analisi del discorso politico nei quotidiani italiani durante la penultima e ultima crisi di governo. I risultati sono pubblicati in alcuni lavori a livello nazionale e internazionale che ho elencato in una breve bibliografia. L'analisi utilizza dati quantitativi classici come il rapporto types/tokens e poi introduce informazioni derivate dal sistema GETARUNS che compie un parsing completo dei testi dal punto di vista sintattico, semantico e pragmatico. I dati riportati nelle tabelle sono derivati dai file di output del sistema. Il sistema produce un file per ogni frase, un file complessivo per l'analisi semantica del testo e un file con la versione verticalizzata del testo analizzato dove ogni parola è accompagnata da una classificazione sintattico-semantica-pragmatica. Il sistema è composto da un parser a reti di transizione aumentate da informazioni di sottocategorizzazione, che costruisce prima i chunks e poi a cascata le strutture a costituenti complesse più alte fino a quella di frase. Questa rappresentazione viene passata a un altro parser che lavora a isole, partendo da ciascun complesso verbale, corrispondente al costituente verbale. Il parser a isole individua la struttura predicato-argomentale, includendo anche gli aggiunti sulla base delle informazioni contenuto in un lessico di sottocategorizzazione per l'italiano costruito in precedenti progetti, contenente circa 50mila entrate verbali e aggettivali a diversi livelli di profondità. Viene utilizzata anche una lista di preferenze di selezione per verbi, nomi e aggettivi ricavata dai treebanks di italiano disponibili e contenente circa 30mila entrate. Riportiamo in Tabella 1. i dati in numeri assoluti.

[1] "Il mio sondaggio aveva già dato la vittoria a Bugliesi – I risultati di una ricerca di un docente di Linguistica" - Il Gazzettino – VeneziaMestre, mercoledì 11.06.2014,p.7.

Candidato	Tokens	Types	Rare words	Frasi	Little Pro	Strutture Proposiz.
Bertinetti	4992	1561	1341	162	200	585
Brugiavini	2841	987	852	91	119	308
Bugliesi	13210	2483	1899	463	541	1232
Cardinaletti	5346	1479	1243	167	159	469
LiCalzi	14376	3120	2516	769	720	1624

Tabella 1. Dati assoluti dei testi analizzati

2. I risultati dell'analisi

Mostriamo in questa sezione i risultati comparati dell'analisi su tutti i livelli linguistici. Commentiamo ogni grafico descrivendo il contenuto in forma verbale senza fornire alcuna valutazione oggettiva, cosa questa che faremo in una sezione finale del lavoro. Questi risultati sono quelli resi pubblici sul forum dei candidati rettore prima dell'elezione. In Fig.1 nell'Appendice, il grafico associato ai dati sintattico-semantici. I dati sono riportati in frequenze relative in modo da poter essere confrontabili, visto che i testi dei programmi sono di lunghezza diversa. Partendo dall'alto il parametro NullSubject misura le frasi semplici a verbo flesso (quindi non quelle a tempo indefinito come le infinitive) che non hanno soggetto espresso. Il secondo parametro misura le frasi che esprimono un punto di vista soggettivo (sono quindi rette da un verbo del tipo di "pensare, credere" ecc.). Il terzo parametro individua le frasi semplici o clausole o proposizioni a polarità negativa, che quindi contengono una negazione a livello verbale (un NON, ma anche MAI ecc.). In questo caso, sono considerate non a polarità negativa le frasi rette da un verbo con significato negativo lessicalizzato accompagnate dalla negazione. Mi riferisco a verbi del tipo di "distruggere, rifiutare, odiare", ecc.

Il quarto parametro misura le frasi non fattive o non fattuali, che cioè non descrivono un fatto avvenuto o che esiste nel mondo. Queste frasi sono rette da verbi di modo irreale (come il condizionale, il congiuntivo, ma anche dal tempo futuro) o sono in forma non dichiarativa, come le domande o le imperative. Come si può evincere dal grafico, Cardinaletti e Bugliesi hanno il maggior numero di frasi non fattive. Invece LiCalzi fa un uso più elevato di frasi a soggetto nullo assieme a Bugliesi, e di frasi soggettive.

Nel secondo grafico (vedi Fig.2 in Appendice) sono analizzati gli aspetti affettivi – questa analisi è chiamata Sentiment Analysis, e contiene anche dati semantici sulla complessità testuale. Sulla base di un lessico specializzato per l'italiano si contano le parole che hanno "prevalentemente" un valore negativo vs. positivo. Il lessico è composto da SentiWordNet (Esuli, Sebastiani, 2006), opportunamente corretto per tutte le parole con valore ambiguo; e contiene un lessico specializzato di circa 70mila entrate prodotto manualmente da me sulla base dell'analisi di testi di giornale per 1 milione di tokens. Le parole con valore neutro non vengono prese in considerazione. Un terzo parametro è quello dell'uso della diatesi passiva, che ha come funzione testuale di permettere la cancellazione dell'agente per far risaltare l'oggetto del verbo e trasformarlo in Argomento principale (o Topic) del discorso. Come si evince dal grafico, il numero maggiore di parole positive è di Brugiavini e Bugliesi, mentre il maggior numero di parole negative è di LiCalzi e Brugiavini. Bugliesi ne utilizza meno di tutti. Per quanto riguarda la forma passiva di nuovo Bugliesi è quello che ne usa di meno, invece Cardinaletti ne usa più di tutti. Per quanto riguarda la complessità, viene riportata la proporzione di proposizioni semantiche per frase, includendo in questo frasi semplici, clausole o complessi predicativi composti da un verbo a tempo indefinito e suoi argomenti. La maggior complessità spetta a Brugiavini e Bertinetti. LiCalzi ha quella più bassa.

Nel terzo grafico (Fig.3 in Appendice) si mostrano dati quantitativi della Vocabulary Richness (VR) in basso, derivati dal conteggio del numero di occorrenze di forme di parola singole chiamate Types, rispetto al totale delle occorrenze chiamate Tokens (queste includono anche la punteggiatura), e indicate dall'abbreviazione TT. La formula in alto invece rappresenta il rapporto che interviene tra i Types e le Rare Words (RW), che sono tutte le forme di parola che ricorrono una volta, due volte e tre volte nel testo, e sono anche chiamate Hapax, Dis e Tris Legomena. I dati rappresentati vedono Brugiavini e Bertinetti come quelli con la più alta ricchezza di vocabolario, e Bugliesi con i valori più bassi. Il rapporto Tokens/Sentence ci dice che LiCalzi ha quello più basso seguito da Bugliesi, mentre gli altri tre testi sono più o meno allo stesso livello.

Per studiare meglio nel dettaglio i concetti che hanno caratterizzato i vari programmi abbiamo

quindi fatto ricorso a una comparazione delle Rank List ricavate dalle liste di frequenza – che non possiamo qui includere per mancanza di spazio. La Rank List è la lista delle parole Types fatta sulla base della loro frequenza. La posizione nella lista indica la rilevanza che la parola assume all'interno del testo. Benché le frequenze assolute siano diverse da testo a testo, la posizione nella rank list permette di valutare le differenze/somiglianze tra testi diversi nella utilizzazione di certe parole chiave.
Tutti i candidati hanno la stessa parola all'inizio della rank list, "ateneo". Anche le posizioni reciproche di "ricerca" e "didattica" e "studenti" sono molto vicine e sono rispettivamente in terza posizione "ricerca" (seconda per Cardinaletti), e in quarta posizione "didattica" (seconda Bertinetti e quinta Bugliesi). Poi le liste si differenziano: la parola "dipartimenti" viene trattata in maniera diversa da LiCalzi che la posiziona molto in alto, mentre Bertinetti la posiziona in basso e in Brugiavini non la si ritrova nelle prime 30. La parola "personale" appare nei testi dei primi tre candidati ma non in quelli di Bertinetti e Brugiavini. Lo stesso dicasi per "valutazione". Invece per quanto riguarda la parola "lavoro" vediamo che essa risulta in alto nella lista dei due candidati Brugiavini e Bertinetti, a differenza di quanto avviene nelle liste degli altri tre candidati dove si trova spostata in basso. Tornerò al contenuto della Rank List più avanti.

3. Calcolo della Correlazione

Infine ho eseguito il calcolo della correlazione tra i vari candidati. Ho utilizzato i vettori delle 11 feature presentate prima, in valori assoluti come parametri di confronto. Nella valutazione, ho considerato solo i casi in cui l'indice R supera 0.998. Il risultato più alto è stato ottenuto dal confronto Brugiavini e Bertinetti, seguono LiCalzi e Bugliesi.

1. ***Brugiavini/Bertinetti***
R = 0.9988378753376379.
2. ***Bugliesi/LiCalzi***
R = 0.9988321771943326.

I valori della ***Cardinaletti*** sono risultati vicini solo a quelli di ***Brugiavini***.

R = 0.9961624024306578.

4. Analisi Semantica Dettagliata di un concetto: PERSONALE

Ho verificato nel dettaglio i dati relativi al concetto PERSONALE che indico in basso. I dati sono limitati ai quattro documenti dei candidati che parlano di PERSONALE in maniera consistente. Abbiamo escluso dal conteggio i due candidati Bertinetti e Brugiavini perché i numeri assoluti nel loro caso sono così esigui rispetto al numero complessivo di TYPES che ricadono nelle cosiddette Rare Words, cioè le parole utilizzate come Hapax, Dis o Trislegomena. Queste parole fanno parte della coda della distribuzione e non contribuiscono a caratterizzare il testo.
Ho contato le volte che la parola viene utilizzata come Nome, come Aggettivo, e come parte della Forma Polirematica Personale_Docente.

	Nome	Aggettivo	Multiword	Totale
LiCalzi	22	4	5	17
Cardin.	11	2	4	7
Bugliesi	37	2	5	32

Tabella 1. Utilizzo del concetto *Personale*

Se si considera quindi che il significato voluto della parola PERSONALE si ottiene solo quando è utilizzata come Nome escludendo le occorrenze dello stesso nome nella forma polirematica, abbiamo Bugliesi primo, seguito da LiCalzi e Cardinaletti.
Nei testi però, si utilizzano descrizioni linguistiche diverse per riferirsi alla stessa entità - persona, organizzazione, località o istituzione. Per quanto riguarda l'uso di coreferenti al concetto PERSONALE abbiamo considerato i due iponimi, PTA e CEL, di cui elenchiamo le seguenti quantità assolute e relative:

- Cardinaletti 8 0.72
- LiCalzi 7 0.32
- Bugliesi 6 0.16

Per ricavare valori relativi, abbiamo fatto la proporzione tra l'uso del riferimento generico PERSONALE e i suoi iponimi. Si conferma l'ordine sulla base dei dati assoluti.

5. Valutazione e Conclusione

Volendo fare una graduatoria complessiva, si può considerare che ciascun parametro possa avere valore positivo o negativo. Nel caso fosse

positivo la persona con la quantità maggiore si vedrà assegnare come ricompensa il valore 5 e gli altri a scalare un valore inferiore di un punto, fino al valore 1. Nel caso invece che il parametro avesse valenza negativa, il candidato con la quantità maggiore riceverà al contrario il punteggio inferiore di 1 e gli altri a scalare un valore superiore di un punto fino a 5. La graduatoria complessiva verrà quindi stilata facendo la somma di tutti i punteggi singoli ottenuti. L'assegnazione della polarità a ciascun parametro segue criteri linguistici e stilistici, ed è la seguente:

1. NullSubject - positive: La maggior quantità di soggetti nulli indica la volontà di creare un testo molto coeso e di non sovraccaricare il riferimento alla stessa entità con forme ripetute o coreferenti.
2. Subjective Props - negative: La maggior quantità di proposizioni che esprimono un contenuto soggettivo indica la tendenza da parte del soggetto di esporre le proprie idee in maniera non oggettiva.
3. Negative Props - negative: Il maggior uso di proposizioni negative, cioè con l'utilizzo della negazione o di avverbi negativi, è un tratto stilistico che non è propositivo ma tende a contrastare quanto affermato o fatto da altri.
4. Nonfactive Props - negative: L'utilizzo di proposizioni non fattive indica la tendenza stilistica ad esporre le proprie idee utilizzando tempi e modi verbali irreali - congiuntivo, condizionale, futuro e tempi indefiniti.
5. Props / Sents - negative: Il rapporto che indica il numero di proposizioni per frase viene considerato in maniera negativa a significare che più è elevato maggiore è la complessità dello stile.
6. Negative Ws - negative: Il numero di parole negative utilizzate in proporzione al numero totale di parole ha un valore negativo.
7. Positive Ws - positive: Il numero di parole positive utilizzate in proporzione al numero totale di parole ha un valore positivo.
8. Passive Diath - negative: Il numero di forme passive utilizzate viene considerato in maniera negativa in quanto oscura l'agente dell'azione descritta.
9. Token / Sents - negative: Il numero di token in rapporto alle frasi espresse viene trattato come fattore negativo di nuovo in riferimento al problema della complessità indotta.
10. Vr - Rw - negative: Questa misura considera la ricchezza di vocabolario sulla base delle cosiddette RareWords, o numero complessivo di Hapax/Dis/Tris Legomena nella Rank List. Maggiori sono le parole uniche o poco frequenti più lo stile è complesso.
11. Vr - Tt - negative: Come sopra, questa volta considerando il numero totale dei Tipi.
L'assegnazione del punteggio sulla base dei criteri indicati definisce la seguente graduatoria finale:

Bugliesi	47
LiCalzi	36
Brugiavini	28
Cardinaletti	27
Bertinetti	27

Tabella 2. Graduatoria finale sulla base degli 11 parametri (vedi Tab. 2.1 in Appendice 2)

Volendo includere anche i punteggi relativi all'uso di PERSONALE e dei suoi iponimi avremo questo risultato complessivo:

Bugliesi	53
LiCalzi	44
Brugiavini	37
Cardinaletti	31
Bertinetti	30

Tabella 3. Graduatoria finale sulla base dei 13 parametri (vedi Tab. 3.1 in Appendice 2)

Utilizzando i parametri come elementi di giudizio per classificare lo stile dei candidati e assegnando una valutazione a parole, si ottengono i due giudizi sottostanti.
1. Bugliesi ha vinto perché ha utilizzato uno stile più coeso, con un vocabolario più semplice, delle strutture sintattiche semplici e dirette, esprimendo i contenuti in maniera concreta e fattuale, parlando a tutti i livelli di parti interessate, docenti e non docenti. Inoltre ha utilizzato meno espressioni e frasi negative e più espressioni positive.
I dati ci dicono anche che il programma di Bugliesi è in forte correlazione con quello di LiCalzi ma non con quello degli altri candidati.
2.Cardinaletti ha scritto un programma che utilizza uno stile poco coeso, con un vocabolario alquanto elaborato, con strutture sintattiche abbastanza più complesse, esprimendo i contenuti in maniera molto meno concreta e molto meno fattuale, parlando a tutti i livelli di parti interessate, docenti e non docenti. Inoltre ha utilizzato poche espressioni e frasi negative e relativamente poche espressioni positive. Infine il programma della Cardinaletti è in buona correlazione con il programma della Brugiavini.

Bibliografia

Delmonte, R., 2013, Extracting Opinion and Factivity from Italian political discourse, in B. Sharp, M. Zock, (eds), Proceedings 10th International Workshop NLPCS, Natural Language Processing and Cognitive Science, 162-176, Marseille.

Delmonte, R., 2012. Predicate Argument Structures for Information Extraction from Dependency Representations: Null Elements are Missing, 2013. in C. Lai, A. Giuliani and G. Semeraro (eds.), DART 2012: Revised and Invited Papers, "Studies in Computational Intelligence", Springer Verlag, 1-25.

Delmonte, R., Daniela Gifu, Rocco Tripodi, 2013. Opinion and Factivity Analysis of Italian political discourse, in R. Basili, F. Sebastiani, G. Semeraro (eds.), Proc. 4th Italian Information Retrieval Workshop, IIR2013, Pisa. CEUR Workshop Proceedings (CEUR-WS.org), http://ceur-ws.org, vol. 964, 88-99.

Delmonte, R., D. Gifu, R. Tripodi, 2012. A Linguistically-Based Analyzer of Print Press Discourse, International Conference on Corpus Linguistics, Saint Petersburg, at http://corpora.phil.spbu.ru/talks2013.

Delmonte R. & Daniela Gîfu, 2013. Opinion and Sentiment Analysis of Italian Print Press, in International Journal Of Advanced Computer And Communication Technology (IJACCT), Vol. 1, Issue. 1, 24-38, http://www.scribd.com/doc/.

Delmonte R. & Vincenzo Pallotta, 2011. Opinion Mining and Sentiment Analysis Need Text Understanding, in Pallotta, V., Soro, A., Vargiu, E. (Eds.), "Advances in Distributed Agent-based Retrieval Tools: Studies in Computational Intelligence, Vol. 361, Springer, 81-96.

Delmonte, R., 2010. Keynote Speaker, OPINION MINING, SUBJECTIVITY and FACTUALITY, ALTA (Australasian Language Technology Association) Workshop, ICT University of Melbourne,http://www.alta.asn.au/events/alta2010 / alta-2010-program.html.

Esuli A, Sebastiani F, 2006. SentiWordNet: A Publicly Available Lexical Resource for Opinion Mining. In: Proceedings of LREC 2006 – 5th Conference on Language Resources and Evaluation.

APPENDICE 1.

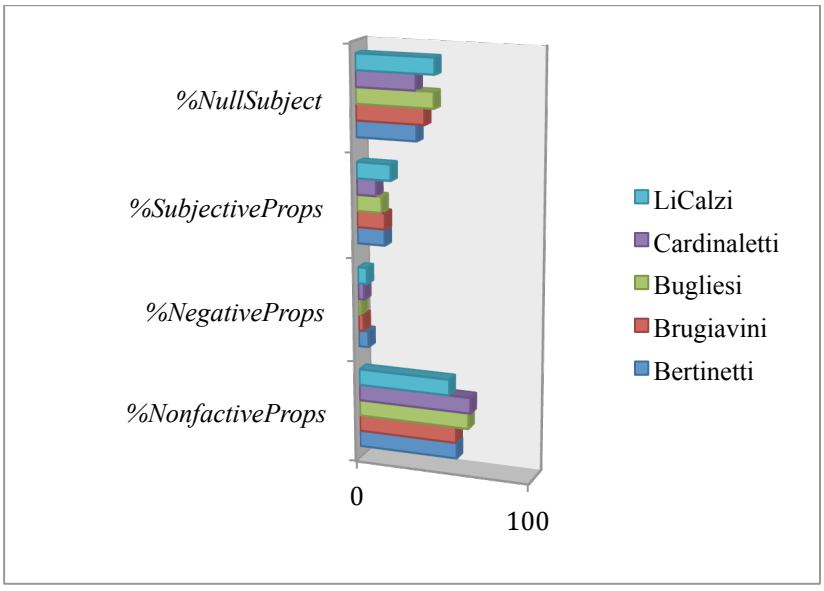

Figura 1. Dati Sintattico-Semantici.

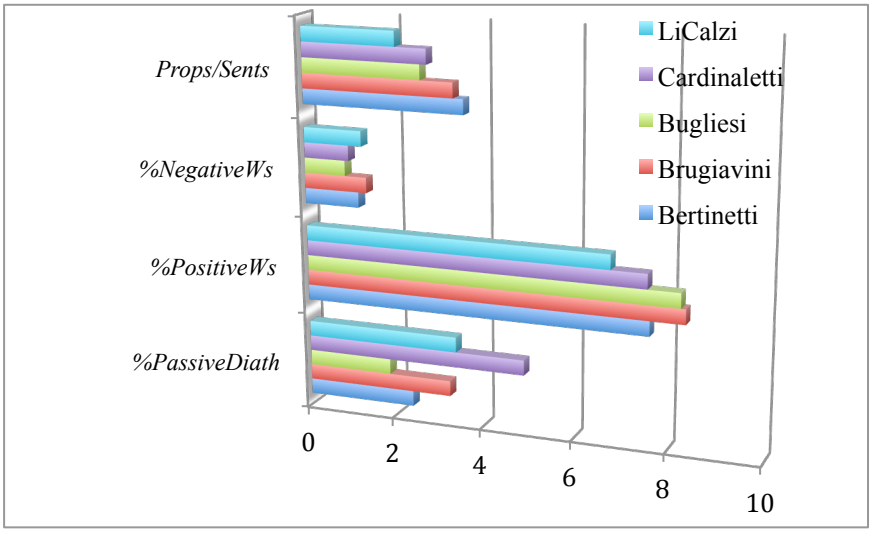

Figura 2. Dati Affettivi e Semantici.

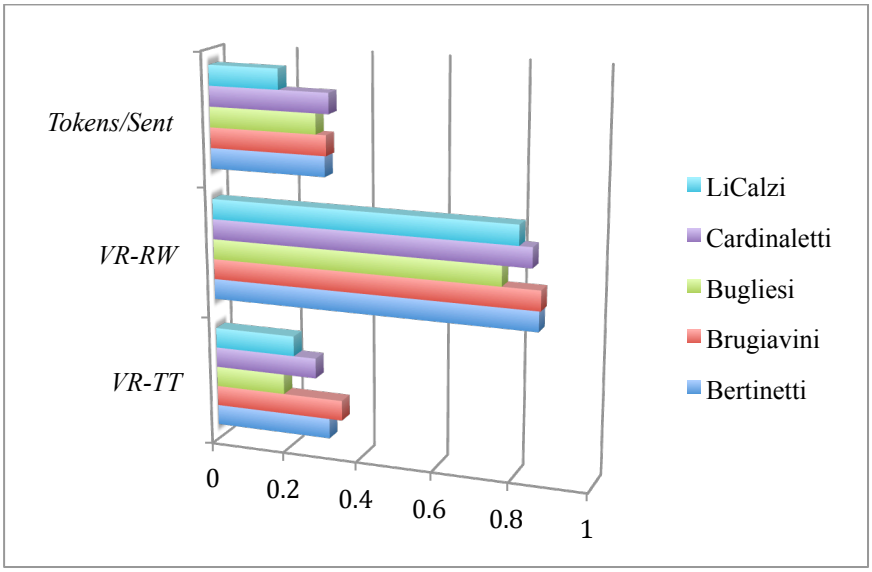

Figura 3. Dati Quantitativi.

APPENDICE 2

	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	Totale
Bugliesi	4	4	5	2	4	5	4	5	4	5	5	47
LiCalzi	5	1	2	5	5	2	1	2	5	4	4	36
Brugiavini	3	2	4	4	2	1	5	3	2	1	1	28
Cardinaletti	2	5	3	1	3	4	2	1	1	3	3	27
Bertinetti	1	3	1	3	1	3	3	4	3	2	2	27

Tabella 2.1 Graduatoria finale sulla base degli 11 parametri.

	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	P13	Totale
Bugliesi	4	4	5	2	4	5	4	5	4	5	5	3	3	53
LiCalzi	5	1	2	5	5	2	1	2	5	4	4	4	4	44
Cardinaletti	2	5	3	1	3	4	2	1	1	3	3	5	5	37
Brugiavini	3	2	4	4	2	1	5	3	2	1	1	2	1	31
Bertinetti	1	3	1	3	1	3	3	4	3	2	2	1	2	30

Tabella 3.1 Graduatoria finale sulla base dei 13 parametri.