Riforma della Pubblica Amministrazione italiana. Text mining e sentiment della consultazione pubblica

Pasquale Pavone[1], Francesca Dolcetti[2], Alessio Canzonetti[3], Barbara D'Amen[4]

[1]Dipartimento FISPPA, Università degli studi di Padova – Italia

[2] Studio RisorseObiettiviStumenti, Roma – Italia

[3]Consorzio MIPA, Roma, Italia

[4]Dipartimento di Scienze Statistiche, Università "La Sapienza" di Roma – Italia

Abstract

Within the context of the public administration reform project, the Italian government, headed by Matteo Renzi, opened to a public consultation. Within the established period of thirty days, civil servants and all citizens had been able to send their opinion regarding the 44 points which constitute the proposed government reform. The site had received 39.343 valid e-mails, then further reduced to 32.443 analyzed by automatic text analysis and applying the tools and techniques of text mining. the paper intends to give an account of the choices made to extract information and classify the e-mails in particular to qualify the sentiment of the texts.

Abstract

Nell'ambito del progetto di riforma della Pubblica amministrazione, il governo italiano presieduto da Matteo Renzi ha lanciato nella primavera del 2014 una consultazione pubblica. Nello spazio di trenta giorni, dipendenti pubblici e tutti i cittadini hanno potuto inviare la propria opinione in merito ai 44 punti costitutivi la proposta di riforma governativa. Sono pervenute 39.343 e-mail valide, poi ulteriormente ridotte a 32.443 analizzate mediante analisi automatica dei testi, applicando strumenti e tecniche del text mining. Questo lavoro intende dar conto dei risultati delle scelte operate per estrarre informazioni e classificare le e-mail in particolare per qualificare il sentiment dei testi.

Key words : text mining, sentiment, espressioni regolari (ER), riforma Pubblica Amministrazione, consultazione pubblica, 44 punti Renzi.

1. Strumenti e tecniche di text mining per la rappresentazione di queste e-mail

Nell'ambito del progetto di riforma della Pubblica amministrazione (PA), il governo italiano presieduto da Matteo Renzi ha lanciato nella primavera del 2014, una consultazione pubblica attraverso l'apertura di una casella di posta elettronica, rivoluzione@governo.it. L'indirizzo e-mail aperto dal Governo aveva l'obiettivo di consentire a tutti i cittadini, quindi non solo ai dipendenti pubblici, di esprimersi liberamente in merito ai 44 punti costitutivi la proposta di riforma governativa della PA, inviando la propria opinione nello spazio di trenta giorni. Alla mezzanotte del 30 maggio 2014, ora di chiusura della consultazione, erano pervenute 39.343 e-mail valide (già depurate all'origine di mail completamente vuote o con solo allegato). A partire da queste, sono state messe da parte 6.900 e-mail rispondenti ad un oggetto che

identificava una petizione, e che avrebbero rappresentato degli outliers per cui nel presente lavoro si analizzano 32.443 e-mail.[1]

Questa innovazione nella partecipazione alla decisione politica ha posto il problema di come analizzare una mole imponente di documenti. Obiettivo finale del lavoro[2] è stato fornire uno studio delle e-mail per creare statistiche sui singoli punti della consultazione e soprattutto selezionare quelle e-mail che maggiormente contenessero proposte sul merito dei punti di riforma proposti, in modo da orientare la lettura diretta delle e-mail stesse da parte dei funzionari del ministero.

Qui si intende dar conto delle strategie di applicazione di strumenti e tecniche di text mining che hanno permesso una classificazione e una categorizzazione automatica dei testi attraverso quelle rappresentazioni capaci di riorganizzare i differenti livelli di contenuti presenti nei testi delle e-mail. Questo ha comportato sia il recupero di informazioni (Information Retrieval -IR) ossia la ricerca e cattura di documenti con caratteristiche attinenti ad alcune *query* create ad hoc secondo criteri di rilevanza o salienza, sia l'estrazione di informazioni (Information Extraction - IE) volte a localizzare, mediante strumenti di evidenziazione (tags) di elementi delle query nei documenti rintracciati (Bolasco, 2013). Ogni rappresentazione è stata ottenuta grazie ad un "modello" espresso sotto forma di interrogazione a tre livelli: a) di parole in quanto tali - staccate dal loro contesto – , b) del contesto locale, ciò che precede la parola e ciò che la segue, c) del documento, nel nostro caso l'unità di analisi dello studio è stata la singola e-mail. Le interrogazioni sono state realizzate con query, sia lessicali che testuali, a differente livello di complessità: dalla ricerca di una lista di parole, alla creazione di classi di parole, uniformi per tipo grammaticale (es. verbi) o per tematica/categoria di interesse (es. professioni, sentiment); tali classi sono poi state messe in relazione fra loro. I modelli sono stati migliorati progressivamente all'aumentare dei dati disponibili. Nel nostro caso l'addestramento dei modelli si è avvalso di quattro aggiornamenti di arrivi di mail (9.000, 18.000, 24.000 e 32.000) nel corso dei trenta giorni della consultazione.

Il lavoro[3] che qui presentiamo ricostruisce i principali passi di una filiera composta di differenti "algoritmi" che nel seguito sono denominati anche modelli, più avanti si proporranno esempi di alcuni di essi che hanno fatto parte della procedura di analisi. In particolare i passi di questa strategia sono stati volti ad estrarre informazioni significative sui contenuti dei testi, classificare le mail secondo criteri e indici di rilevanza, qualificare il sentiment dei testi e ricavare statistiche sulle caratteristiche degli scriventi. L'intera analisi è stata effettuata utilizzando il software Taltac2, la cui tipicità è quella di permettere un'alternanza fra momenti di analisi lessicale e momenti di analisi testuale (Bolasco, 2010).

[1] L'insieme di tali e-mail ammontano a un testo di diciottomila pagine.

[2] Il lavoro è stato condotto dal gruppo di ricerca coordinato dal Prof. Sergio Bolasco presso il Dipartimento MEMOTEF della Facoltà di Economia dell'Università La Sapienza di Roma. I risultati di questo lavoro sono stati oggetto di un report interno, 1.03 del 16 giugno 2014, per il Dipartimento della funzione pubblica, struttura della Presidenza del Consiglio dei Ministri.

[3] Nonostante questo lavoro sia un prodotto condiviso degli autori, si può attribuire ad Alessio Canzonetti la stesura del paragrafo 2, a Barbara D'Amen la stesura del paragrafo 3.1., a Francesca Dolcetti la stesura dei paragrafi 1, 3.2, 3.3 e conclusioni, a Pasquale Pavone la stesura del paragrafo 4, a Francesca Dolcetti e Pasquale Pavone hanno curato assieme la stesura del paragrafo 5."

2. Identificazione delle caratteristiche personali dei rispondenti nel corpus

Il corpus in analisi risulta composto da 32.443 e-mail considerate valide. Sono state eliminate a priori le e-mail vuote, senza contenuto (o che possedevano soltanto un allegato, senza quindi contenuto testuale alcuno), e quell' insieme di e-mail di protesta/provocazione politica, identificate come una petizione, aventi tutte il medesimo oggetto. L'ampiezza del corpus è pari a 9,4 milioni di occorrenze di parole, con un vocabolario (inteso come varietà di linguaggio rilevato) di oltre 200.000 parole diverse, incluse decine di migliaia di indirizzi e-mail intesi come unica parola, del tipo *nome@dominio*.

L'aver condotto la consultazione via e-mail ha determinato la necessità di estrarre alcune caratteristiche personali dei rispondenti direttamente dal contenuto testuale delle stesse. I rispondenti non erano infatti tenuti a compilare alcun form di registrazione/identificazione, né è stato loro chiesto di produrre le loro e-mail conformandosi ad un particolare format che avrebbe potuto facilitare l'individuazione di certi item (genere, professione, età, residenza ecc.). Sono state quindi messe a punto una serie di procedure di ETL (Canzonetti et al., 2006) volte ad individuare l'eventuale presenza di certe caratteristiche ed estrarne il contenuto.

Per l'identificazione del **genere** si sono utilizzati due elenchi di nomi maschili e femminili, andandoli a cercare nell'indirizzo e-mail, nella presentazione o nei saluti (quindi all'inizio o alla fine del testo della e-mail). Un'esempio di espressione di ricerca è la seguente:

```
Buongiorno, mi chiamo {NOME}
```

Per la **localizzazione geografica**, attraverso una procedura di normalizzazione del testo basata anche su liste di riconoscimento, sono stati individuati tutti i toponimi presenti. Sono stati quindi considerati tutti quelli relativi a nomi di comuni italiani che però non fossero presenti nelle strutture sintattiche proprie di un indirizzo ("Via Roma", "Piazza Bologna", "Corso Trieste" ecc.). Le e-mail per cui è stato possibile rilevare il genere del rispondente sono state il 73,7%, mentre molte meno (32,4%) sono state invece le e-mail identificate nella zona geografica.

L'accertamento delle **caratteristiche professionali** e l'**afferenza ad un particolare ente pubblico** ha riguardato solo una parte di e-mail, ovvero di coloro che hanno liberamente espresso la propria posizione in tal senso. L'ampiezza numerica dei risultati, superiore a 4.000 casi, per quanto ristretta rispetto alle e-mail disponibili ha consentito, comunque, di avere già una sufficiente stima della situazione. Lo studio di tali caratteristiche non era privo di oggettive difficoltà. Per fare un esempio: l'informazione, quando è completa, si presenta in questa forma: *"buongiorno, sono un dipendente statale lavoro come tecnico manutentore presso un'azienda ospedaliera"* (dichiarazione di qualifica, ruolo ed ente di afferenza). In realtà, spesso erano presenti solo una o due di queste componenti. Pertanto, l'individuazione dell'informazione è avvenuta attraverso la messa a punto di un modello sintattico che replica, in astratto, le principali modalità in cui viene dichiarata la propria situazione lavorativa ed eventualmente l'ente/azienda di appartenenza, ed è così riassumibile:

```
incipit +infisso + professione
```

in cui l'**incipit** è dato da espressioni del tipo *"mi chiamo*", *"sono"* o *"lavoro"*, l'**infisso** è del tipo *"lavoro"*,*"presso"* ecc., e l'**elemento variabile** è costituito da una **lista di professioni**. Il modello così implementato ha dato luogo ad una serie di interrogazioni sul testo che ha permesso di intercettare citazioni del tipo: *"mi chiamo (nome) Funzionario Amministrativo*

del Comune di Pordenone..."; *"lavoro come dirigente amministrativo di un ente pubblico..."*; *"sono un dipendente pubblico dal 1983 e intendo formulare..."*; *"sono un dipendente pubblico uno di quelli attualmente nell'occhio del ciclone infatti lavoro presso la Motorizzazione di Bari..."* e molte altre. Come si può notare, non sempre la dichiarazione della propria professione è puntuale e precisa, né viene sempre dichiarata la propria appartenenza. Citazioni di questo tipo sono state rilevate in 4.196 e-mail, che sono state quindi estratte e sottoposte a ulteriore analisi volta ad attribuire ad ognuna una categorizzazione circa la professione e l'ente di appartenenza.

In questa prima fase di cattura, il modello è stato impostato "aperto" verso destra, in maniera da raccogliere anche un certo numero di parole successive che, generalmente, contengono l'ente di appartenenza. In tal modo non è stato necessario impostare una lista di enti, circostanza molto laboriosa a causa del loro elevato numero e, soprattutto, della molteplicità di forme in cui uno stesso ente può essere citato. Nel totale delle e-mail analizzate, le qualifiche maggiormente ricorrenti, oltre alla generica dizione di "dipendente", riguardano posizioni tipiche dell'impiego pubblico (Segretario, dirigente, responsabile, funzionario ecc.), seguite da parole da un lato più specifiche (insegnante, medico, avvocato, ricercatore ecc.), dall'altro semanticamente ambigue (informatico, tecnico, contabile, legale ecc.). Le occorrenze di queste ultime sono pertanto sovrastimate, poiché ricomprendono casi in cui non si sta citando una professione, bensì l'aggettivo. In totale le professioni accertate sono 179 per complessive 44.351 citazioni.

Tra gli enti maggiormente citati si annoverano le varie declinazioni di ente locale – Comune, Provincia, Regione , Comunità (montana), Unione (di comuni). Altre citazioni molto frequenti: Istituto (scolastico), Ministero, Agenzia (in varie declinazioni: Lavoro, Dogane, Entrate, Demanio ecc.).

Inoltre, come anticipato, dall'intero insieme delle e-mail analizzate è stato estratto un **sottoinsieme di 4.196 citazioni** contenenti la dichiarazione della professione svolta ed eventualmente la propria afferenza a un ente. Questo sub-corpus è stato analizzato in maniera indipendente ed approfondita per produrre statistiche riguardanti la professione svolta dai rispondenti e la loro organizzazione di appartenenza. In gran parte dei casi, la dichiarazione di professione esercitata è generica (dipendente, 1.276 casi pari al 30,4%, in gran parte si intende "pubblico") o non precisata (*"lavoro presso il Comune di ..."*, 485 casi, 11,6%). Si incontrano anche varie funzioni superiori (funzionario, dirigente, responsabile ecc.) e quelle più specifiche dei settori non amministrativi (insegnanti, medici, tecnici, ingegneri) o di rango inferiore (impiegato, assistente ecc.). Gli enti associati alle professioni dichiarate risultano essere molto meno, 1.715, poiché l'afferenza spesso non è citata oppure viene, per così dire, "incorporata" nella definizione di "dipendente pubblico", vale a dire di una generica amministrazione pubblica. Gli enti locali sono i più presenti (38%: Comune, Provincia, Regione, enti associativi ecc.) seguiti dalla Scuola (10,6% a cui si devono aggiungere la maggior parte delle 68 varietà della voce Istituto). Rilevante anche la presenza di: ministeri, Agenzia delle entrate, Inps e Asl. L'ente di appartenenza è stato quindi rilevato per 1.769 e-mail, ovvero nel 42% dei casi

3. Il lessico degli scriventi: contenuti e sentiment

3.1. Le parole tema e le parole chiave

L'analisi del lessico degli scriventi ha considerato sia le "parole tema", ovvero le parole più frequenti che compongono il vocabolario, sia le "parole chiave" o peculiari, ritenute tali poiché, nel corpus in analisi, presentano un uso diverso rispetto a quello che mediamente se ne fa nel linguaggio comune[4] (Bolasco, 2013, p. 135). Considerando le parole tema, la parola più frequente è l'oggetto stesso del corpus, nella fattispecie "pubblica amministrazione", rinvenibile anche con le espressioni "PA" e "amministrazione pubblica", per un totale di 34.070 occorrenze. Segue come più frequente la parola "lavoro"(18.960 occorrenze) già al netto di espressioni quali "buon lavoro" (3.588), "posto/i di lavoro" (1.210) e < contratto /datore /mondo /orario /rapporto di lavoro ecc.>. Ciò riprova che il tema centrale in questa consultazione è il lavoro; questo presenta quasi il doppio delle occorrenze rispetto alle altre parole più frequenti - dipendenti, riforma, personale - tutte con occorrenze superiori a 10.000. L'analisi delle parole peculiari rivela aspetti particolarmente tipici; tuttavia, prima di procedere nella presentazione dei risultati, è bene specificare che questa fase di analisi è stata condotta separatamente per due distinti sottoinsiemi: a) email che presentano riferimenti puntuali ed espliciti ad almeno uno dei 44 punti della proposta di Riforma della PA (pari a 8.331); b) email che non hanno esplicitamente citato i suddetti punti, pur trattando probabilmente argomenti relativi alla riforma (pari a 12.338)[5]. Si è ritenuto opportuno fare questa distinzione per "depurare" i risultati dell'analisi dalla presenza delle parole utilizzate nella lettera relativa ai 44 punti della riforma. Sono stati pertanto definiti due sottoinsiemi di email sulla base degli algoritmi di classificazione automatica esposti nel paragrafo 4.

Nel primo sottoinsieme emerge la centralità dei temi contenuti nella riforma della PA, testimoniata dalla significativa presenza di parole quali "abolizione" e "dirigenza", che si riferiscono al punto 10 della lettera, cui seguono le parole "Motorizzazione" e "accorpamento", che riguardano il punto 27. Strettamente collegate ad altri punti della riforma sono le parole "Camera", che quasi sempre si riferisce alla Camera di Commercio, "enti", "digitalizzazione" e "semplificazione". Emerge poi il termine "contrarietà" che esprime il punto di vista critico del rispondente. L'analisi degli aggettivi sovrautilizzati conferma l'aderenza ai temi della riforma con forme grafiche come "camerale" (ente), "camerali" (enti) e "Camerale" (Ente, Sistema) correlabili al punto 29 della lettera di Renzi relativo all'abolizione dell'iscrizione delle imprese alle Camere di Commercio. Come vedremo su queste risultanze vi è una forte incidenza delle email "sollecitate" (cfr. par. 4.1). Inoltre figurano aggettivi come "dirigenziale/i", seguiti da altri direttamente connessi al tema del lavoro e della carriera ("idonei", "concorsuali", "professionali"). L'analisi dei sostantivi originali, mette in evidenza termini legati alle funzioni e agli attori coinvolti, più o meno direttamente, nella riforma e comunque riferiti al settore della PA; tra questi, oltre alla "Confindustria", compaiono "avvocature", "vicedirigenza", "capo-dipartimento". Seguono poi "metrologia" (legale), "interoperabilità" e "reingegnerizzazione", riguardanti le procedure

[4] Per l'individuazione delle parole chiave, che costituiscono il linguaggio peculiare, si calcola un indice di contrasto d'uso che pone a confronto la frequenza relativa che assume una parola nel vocabolario oggetto di analisi e la frequenza relativa che la stessa parola assume in un lessico di frequenza scelto come modello di riferimento; quest'ultimo, nella presente analisi, è rappresentato dal linguaggio dei giornali ("Linguaggio comune – Fg con uso>10 (Rep90)") disponibile nelle risorse statistico-linguistiche di TaLTaC2. Grazie al criterio del linguaggio peculiare è possibile liberarsi dalla dipendenza della sola frequenza per soppesare l'importanza di un termine, prediligendo, invece, come criterio di selezione, l'indice di contrasto d'uso.

[5] Da questo sub-corpus, successivamente, sono state estratte 925 email con un chiaro riferimento a termini della riforma, misurabili rispetto alle tre grandi aree tematiche della riforma piuttosto che ai singoli punti. L'estrazione è avvenuta applicando un appropriato modello di text mining basato sull'indice tfidf (cfr. par. 4.3).

informatiche della PA, "fiduciarietà" e "premialità", maggiormente connessi con le modalità di gestione dei ruoli professionali. Gli aggettivi originali sono ancora riferiti ai temi del lavoro e della progressione di carriera con termini come "apicale/i", "nominabili", "preselettiva/e" (riferito alle prove), "ordinistiche" (riferito alle professioni), seguiti da aggettivi come "metrici" e "interoperabili", riferiti rispettivamente alla metrologia legale e ai processi di informatizzazione della PA.

Le mail appartenenti al secondo sottoinsieme sono state definite "Storie di vita" poiché, alla luce dei processi di text mining applicati, non mostrano riferimenti significativi e diretti ai temi della riforma, ma ne propongono invece una narrazione di tipo personale (11.413 email). L'analisi dei sostantivi sovra-utilizzati mostra la prevalenza del termine "suggerimenti" che testimonia la volontà di partecipazione degli scriventi. Pur trattandosi di testi organizzati secondo la libera narrazione individuale, il lessico peculiare mostra la prevalenza netta dei temi del lavoro e della carriera, con la presenza di parole come: "precariato", "meritocrazia", "mansioni", "servizio", "competenze", "progressione" (di carriera), "indizione" (di concorsi), "turnover" e "amministrazioni". Alla luce di queste considerazioni non sembra casuale la rilevanza, in termini di scarto, che assume il termine "fannulloni" il quale, nei testi analizzati, rimanda a due questioni cruciali: quella del merito e quella della reputazione personale nel contesto lavorativo. Anche gli aggettivi sovra-utilizzati attestano la centralità assunta dai ruoli professionali e dalla possibilità di un loro sviluppo, con termini come "idoneo", "concorsuale/i", "lavorativo/a/e", "professionali", "dirigenziale/i".

Dall'analisi sembra, quindi, che ciò che motiva a rispondere e a proporre suggerimenti siano storie individuali, aventi a oggetto l'ambito professionale, sollecitate dalla proposta di riforma che, in questo caso, diventa l'occasione per dare voce al clima interno della PA. Fra i sostantivi originali di particolare rilevanza sono "governance", "TASI" e "spread", cui si accompagna la presenza di termini più "colloquiali", come "euri" e "bamboccioni". L'analisi degli aggettivi originali è coerente con quanto mostrato dai sostantivi, con la presenza significativa dei temi del lavoro e della progressione di carriera, riconducibile agli aggettivi "apicale/i", "nominabili", "preselettiva/e", "incapienti". Infine, a partire da un approfondimento di analisi delle concordanze sui lessemi <precar*> e <pension*> emerge una narrazione di vicissitudini dovute, tra le altre, alle recenti norme della riforma Fornero, di cui si chiede l'abolizione. Questo aspetto in particolare coinvolge circa il 25% delle mail di questo gruppo.

3.2. Il sentiment dei verbi: dalla valutazione alle proposte

Un'area di lavoro specifica ha riguardato i verbi, quali veicoli della rappresentazione di azioni entro le e-mail. I verbi sono stati considerati a livello di lemmi, sommando tutte le parole che nelle e-mail li hanno declinati. Come per sostantivi ed aggettivi, la prima esplorazione ha utilizzato il criterio dell'estrazione del linguaggio peculiare con una selezione iniziale dei primi 150 verbi ordinati per peculiarità. Dopo aver preliminarmente esplorato i casi di ambiguità attraverso uno studio delle concordanze, è stata realizzata un'analisi di merito per estrarre dai verbi peculiari, calcolati sull'intero corpus, una tonalità generale in termini di positività o negatività dell'azione e dell'atteggiamento (Bolasco e Della Ratta, 2004). Le liste risultanti hanno fatto emergere due questioni rilevanti, che hanno spinto verso un'implementazione del criterio POS-NEG: a) la comune distinzione positivo - negativo assume qui un significato nuovo vista la valenza positiva di alcuni verbi tradizionalmente intesi come negativi (*eliminare, abolire, rottamare, abrogare*), coerenti con la radicalità della

proposta; b) verbi positivi con una peculiarità molto alta, come *augurare*, *sperare, ringraziare,* rappresentavano solo in minima parte il modo con cui si poteva capire in che modo si stava orientando la relazione con la riforma da parte degli scriventi, questi verbi nonostante la loro alta peculiarità si stavano rivelando troppo poco selettivi. Questo ha spinto verso la ricerca di criteri che affinassero ulteriormente la categorizzazione esplorando il vissuto emozionale in quello specifico contesto (Battisti e Dolcetti, 2012). Per i verbi negativi è stato utile il criterio del cambiamento-sovvertimento dello *status quo* capace di cogliere lo spirito di una riforma posta come radicale, e quindi differenziare verbi solo "apparentemente" negativi, ma che stavano assumendo un'accezione positiva, rispetto a quelli cui viene associato comprensibilmente un significato negativo come ad esempio *licenziare, sanzionare, sobbarcare, cestinare, vessare.* La forte peculiarità dei primi, a valenza positiva, sembra indice di un'adesione alle parole d'ordine lanciate dalla consultazione e che interpretano una spinta a fare piazza pulita, a dare un nuovo inizio, dunque un'adesione alla proposta.

Abbiamo poi esteso ai verbi positivi un approccio tratto dagli studi che Osgood, Suci e Tannenbaum (1957) hanno fatto su coppie di aggettivi con l'intento di evidenziarne possibili "strutture cognitive latenti". Gli autori individuarono tre diversi fattori attributivi, soggiacenti al processo di significazione linguistica, e costitutivi dell'atteggiamento soggettivo rispetto ad un oggetto d'indagine: *valutazione* (positivo-negativo), *potenza* (forte-debole), *attività* (attivo-passivo). Tradizionalmente positivo e negativo attengono al primo fattore, quello valutativo, ma poiché in questo lavoro l'intento rilevante era raccogliere suggerimenti e proposte, è sembrato che fosse utile integrare la scelta dei verbi per fondarla anche sugli altri due fattori. In particolare è risultato interessante il fattore *attività* basandoci sul quale l'attenzione si è concentrata sui verbi capaci di rappresentare proposte nell'area organizzativa, determinante per gli scopi della consultazione. Questo ha permesso di differenziare delle categorie, entro i verbi peculiari positivi, che davano spazio all'innovazione in senso costruttivo (*creare*, *innovare*, *rivoluzionare*, *attivare*, *apportare*, *inserire*), a rendere migliore l'esistente (*valorizzare*, *riformare*, *ottimizzare*, *premiare*, *migliorare*, *velocizzare*, *agevolare*, *riorganizzare*, *apprezzare*), a fornire strumenti per il cambiamento (*sopperire*, *garantire*, *supportare*, *dotare*, *motivare*, *responsabilizzare*, *aggiornare*, *accudire*, *aiutare*, *ricevere*), al desiderio di obiettivi (*riuscire*, *guadagnare*, *conseguire*, *risolvere*, *progredire*). Il punto di arrivo dell'ampliamento della categorizzazione oltre il POS-NEG, è stato integrare con una nuova area di verbi, orientanti ad azioni propositive, la spinta radicale rappresentata da verbi ad alta peculiarità, quelli apparentemente negativi ma positivi in cui il motore primo era l'abbandono del vecchio.

3.3. Il sentiment delle posizioni critiche

Lo studio del sentiment nei verbi si è focalizzato anche sull'espressione delle posizioni critiche. Si sono infatti colte espressioni di disagio, più o meno forte, soprattutto quando nei messaggi non si commentano i 44 punti con proposte puntuali, ovvero quando alla critica non si accompagna un contributo per il miglioramento, in altri termini quando si è riusciti meno a stare sugli obiettivi della consultazione. Aspetti di critica e lamentela, fino all'insulto vero e proprio[6], sono stati testati attraverso una lista di verbi, restringendo la ricerca a quelli più

[6] Un termine che più di altri segna non tanto l'insulto quanto lo sdegno è il lessema <vergogn*> declinato nel sostantivo VERGOGNA (spesso gridato!), nell'aggettivo vergognoso, e nel verbo vergognarsi, in una varietà di

capaci di esprimere le difficoltà di chi stava scrivendo. Verbi come *calpestare*, *affossare*, *disfare*, *disperdere*, *mortificare*, *frustrare* per dirne alcuni, sono risultati fra i più capaci di rappresentare insoddisfazione verso il cambiamento. Le mail che li contengono parlano delle difficoltà per le proprie condizioni di lavoro e della PA. Chi si sbilancia sulla critica non propone un lamento tout court, non approfitta dell'ascolto esclusivamente per "vuotare il sacco" ma più spesso si propone di condividere i "mali" della PA assieme alla propria impotenza e amarezza. Messaggi di chi soprattutto teme di non trovare spazio dentro il cambiamento o che segnalano trappole, correttivi, occasioni di competenze sprecate di cui forse tenere conto in un processo di trasformazione. E' una testimonianza più di incomprensione e timore che di diretto antagonismo.

4. Categorizzazione delle e-mail

La categorizzazione delle e-mail è avvenuta in modo diverso in base ad ampiezza[7] e contenuto delle e-mail ricevute. Ci si è lavorato all'interno del DB Documenti di Taltac2. Nei successivi sotto-paragrafi verranno esposte alcune delle procedure adottate.

4.1. E-mail riferite a petizioni

Dopo una prima esplorazione del corpus, effettuato tramite un'analisi delle concordanze, sono stati individuati blocchi di e-mail identiche, successivamente ricondotte alla presenza di specifiche petizioni o sottoscrizione di appelli provenienti da organizzazioni e/o associazioni di vario tipo. Queste e-mail sollecitate hanno trovato differente risposta, sia per l'autorevolezza della fonte (un sindacato o un'organizzazione di categoria piuttosto che un singolo cittadino), sia per la compattezza dei rispondenti (senso di appartenenza alla categoria o alla funzione in via di abolizione o all'ente che fra le proposte di riforma verrebbe accorpato o soppresso). Di conseguenza i corrispondenti blocchi di e-mail hanno diversa consistenza che può essere valutata comunque come "peso" o importanza data all'argomento, in particolare se quest'ultimo è riferito a uno dei 44 punti della riforma. E' il caso dei punti 13 (abolizione del Segretario comunale), 26 (una sola Scuola Nazionale della Pubblica Amministrazione), 27 (accorpamento di Aci, Pra e Motorizzazione) e 29 (eliminazione dell'iscrizione alle CCIAA). I principali ulteriori blocchi di e-mail riguardano invece le sollecitazioni intorno argomenti non contemplati fra i 44 punti di riforma. Uno dei principali fra questi è quello che è stato definito come "il 45mo punto", promosso dalla CGIL, e riferito allo "sblocco o rinnovo dei contratti". Gli altri punti sollecitati tramite una petizione riguardano i seguenti argomenti: "contro la pubblicità dei giochi di azzardo"; "a favore del software libero nella PA"; "riguardante il personale precario dei Centri per l'impiego, della Formazione e delle Province"; "a favore della Siria". Complessivamente le e-mail categorizzate sulla base dell'appartenenza a delle petizioni sono state 11.774 di cui 8.216 e-mail sono pertinenti con i punti della proposta di riforma.

4.2. Le e-mail riconducibili ai punti della riforma

forme diverse, per un totale di circa 800 occorrenze, sparse in centinaia di mail (2% del totale); ma solo in 80 di queste, si trova <vergogn*> almeno 2 volte.

[7] L'ampiezza dei testi delle 32.443 e-mail in analisi è risultata assai variabile, essendo compresa da un minimo di 10 ad un massimo di 6.259 parole.

In una seconda fase dopo aver escluso le e-mail già definite come frutto di petizioni si è proceduto a categorizzare le restanti 20.669 e-mail attraverso la definizione di un modello lessico testuale in grado di categorizzare automaticamente le e-mail rispetto ai punti della riforma. La categorizzazione delle singole unità di contesto rispetto ai punti della riforma è avvenuta mediante la definizione di opportune query testuali consistenti in *espressioni regolari* (ER). Le singole ER sono state definite sulla base dell'ipotesi che all'interno del testo delle e-mail lo scrivente facesse un chiaro riferimento a ciascun punto della riforma prima di esporre la sua opinione in merito. Nella fase di training del modello lessico-testuale (Bolasco, Pavone 2010) si è evidenziato che i riferimenti ai punti della riforma potevano essere di vario tipo. Pertanto le ER sono state costruite in modo da ricercare sia il riferimento generico ai singoli punti, sia il riferimento testuale specifico definito dagli elementi principali dei titoli dei punti della riforma o alcuni sinonimi. Nella fase esplorativa di training sono stati messi a punto i componenti di base del modello, generando delle annotazioni lessicali alle varianti grafiche della parola "*punto*" e definendo alcuni dei più comuni sinonimi degli oggetti delle riforme. Ogni espressione regolare è stata definita in modo di poter identificare i riferimenti diretti ai singoli punti della riforma e di conseguenza categorizzare il documento in cui tali riferimenti sono presenti. Ad esempio per ricercare i riferimenti al *punto 1*[8] della riforma e categorizzare di conseguenza le unità di contesto la ER è stata definita nel seguente modo:

"CATSEM(PUNTO) LAG5 CATSEM(1)" OR "tratteniment? LAG2 servizio" OR "staffetta generazionale" OR "ricambio generazionale"

Allo stesso modo sono state definite le altre ER generando un piano di lavoro testuale costituito da 44 ER, ciascuna delle quali è composta da più elementi di ricerca in OR. L'applicazione del modello ha permesso di categorizzare 8.331 e-mail, di cui 6.780 e-mail riferite esclusivamente ad un punto della riforma e 1.551 e-mail con riferimenti a più di un punto della riforma. A partire da queste categorizzazioni è stato possibile formulare delle statistiche e delle graduatorie relative alle e-mail della riforma "punto per punto". Ai primi posti nella graduatoria dei punti più commentati si trovano i temi che si riconducono ad alcune petizioni quali il 29 (eliminazione iscrizione alle CCIAA), il 13 (abolizione Segretario Comunale), 27 (accorpamento Aci, Pra), mentre spicca rispetto a tutti i restanti con un importante peso il "punto 1" relativo alla "staffetta generazionale", seguito dal "punto 2" sulle modifiche della mobilità. Da questa categorizzazione emergono di fatto 3 aree di attenzione: in primo luogo si trovano le e-mail riferite alle condizioni di vita lavorativa del funzionario; successivamente ci sono i punti relativi alla dirigenza; infine minore attenzione è dedicata alla digitalizzazione e alla semplificazione dei servizi al cittadino.

4.3. Riforma a grandi linee e storie di vita

Dopo le prime due fasi di classificazione restano un totale di 12.338 e-mail in cui non compare un riferimento specifico ai punti della riforma sulla base dei criteri finora definiti. Al fine di categorizzare ulteriormente le e-mail viene definito un secondo modello che attraverso l'utilizzo dell'indice TFIDF, sia in grado di ricondurre le e-mail rispetto alle 3 aree di attenzione rilevate nel precedente paragrafo. La sua applicazione ha permesso di categorizzare ulteriori 925 e-mail (gruppo di e-mail pertinenti che trattano la "riforma a grandi linee") che contengono certamente riferimenti pertinenti ai contenuti della riforma, anche se non specifici

[8] Abrogazione del trattenimento in servizio, sono oltre 10.000 posti in più per giovani nella PA a costo zero.

ai singoli punti. L'ultimo gruppo rimanente di 11.413 e-mail, è stato definito come "Storie di vita" (cfr. par. 3.1). Si tratta di e-mail "pertinenti fra virgolette" dove la proposta di riforma diventa l'occasione per dare spazio anche a quelle voci di corridoio che accompagnano il clima interno della PA.

5. Il Sentiment della Riforma

Il sentiment è stato trattato con Taltac2, costruendo liste personalizzate, inoltre per il positivo-negativo si è lavorato con una risorsa interna al software Taltac per caratterizzare le unità di contesto attraverso dei tag semantici. Dopo aver esplorato, categorizzato e classificato le e-mail con i commenti alla riforma, il lavoro del gruppo di ricerca si è centrato nel fornire delle misure di *sentiment* costruite *ad hoc* per il Corpus in analisi. Il *sentiment* delle e-mail è stato misurato secondo tre punti di vista: A) la presenza di qualificazioni positive e negative in termini di aggettivi nel testo; B) la presenza di "espressioni di accordo" *vs* "disaccordo" rispetto ai punti della riforma; C) la presenza di forme verbali al condizionale, come indicatore di un atteggiamento partecipativo e costruttivo (proposte, suggerimenti o altro). Questi tre criteri sono stati misurati sia nell'insieme del corpus, sia nei diversi gruppi di e-mail.

5.1. Il Sentiment positivo / negativo

Il rapporto fra aggetti positivi e negativi è la modalità più classica di misurare la tonalità di un testo. A partire dal confronto con un dizionario di riferimento composto da 6.000 aggettivi, sono stati estratti nel corpus 2.013 termini positivi e 1.782 termini negativi. In generale un testo si considera con una tonalità negativa se il rapporto fra le occorrenze POS/NEG, secondo i soli aggettivi qualificativi, è al di sotto del valore soglia di 2,5. L'insieme complessivo delle e-mail ha mostrato un *Sentiment* decisamente positivo in quanto questo rapporto è pari a 3,9[9]. Gli aggettivi più frequenti fra quelli che connotano un'opinione negativa, sono *vecchi/a/o* associati a "ruoli" (*vecchi dirigenti*, *vecchi burocrati*, *vecchi baroni*, *vecchi parrucconi*), *inutili/e* (che connota principalmente enti o adempimenti considerati superflui, spesso identificati come *dispendiosi*), *gravi*, *passati*, e via dicendo. Passando agli aggettivi positivi, c'è reciprocità con i significati espressi da quelli negativi: i più frequenti, infatti, esprimono la novità (*nuove/o/i*) associata tendenzialmente all'introduzione di innovazioni, intese sia come nuove pratiche o nuove opportunità lavorative (*nuovi stimoli*, *nuovi servizi, nuovi concorsi*, *nuovi bandi di concorso*). Seguono termini quali *utile, economico* (spesso, utilizzati con riferimento al concetto di efficienza) ecc.

5.2. Il parere favorevole vs contrario

Il parere favorevole o contrario rispetto ai punti della riforma è stato misurato sulla base di termini individuati dalla lettura di un largo campione di testi. In tal modo è stato possibile delineare in maniera certa le forme che indicano tale tipo di sentiment, attraverso parole ed espressioni come: *ottimo, approvo, va bene, benissimo, condivido*, oppure: *non sono d'accordo, disappunto, pericoloso, errore*.

5.3. La propositività nelle e-mail

Dopo una prima esplorazione, fra i rappresentanti del Governo ed il gruppo di ricerca. è stato concordato che un punto rilevante del lavoro fosse il contenuto propositivo presente nei testi;

[9] Il valore è calcolato a partire da 139.450 occorrenze di aggettivi positivi contro 35.920 negativi.

si è quindi convenuto di selezionare un ristretto gruppo di e-mail più propositive per una lettura diretta delle proposte. Per rintracciare un atteggiamento costruttivo e partecipativo, si è scelto di misurare nei documenti la presenza delle forme verbali al condizionale.
Si è osservato, infatti, che tali forme coniugate nella prima persona singolare evidenziano le proposte, i suggerimenti, le osservazioni, le riflessioni o i consigli degli scriventi (*proporrei, farei, eliminerei, aggiungerei*…); allo stesso tempo se coniugate nella terza persona singolare o plurale stanno ad indicare le motivazioni date dalle persone alle loro proposte (*permetterebbero, consentirebbe, faciliterebbe, aiuterebbero*…).
Per quanto riguarda il tipo B, il 12,9% contengono un'espressione di sentiment Contrario e il 17,7% un termine Favorevole; quanto al tipo C, invece solo il 15,5% di e-mail contengono un termine Propositivo. La propositività risulta più elevata nelle e-mail dei gruppi 1) e 2) che argomentano la riforma (dal 23% fino al 33%) e si azzera nelle petizioni (2%). In queste ultime, il parere è favorevole (17%), e non potrebbe essere altrimenti perché è stato appunto richiesto (chi scrive aderisce ad una sollecitazione), mentre è assente un parere contrario (4%) e la qualificazione con aggettivi è inferiore alla media. Viceversa, nei gruppi di e-mail (1 e 2) che commentano la riforma, le percentuali del sentiment di tipo B e C sono notevolmente superiori alla media, a riprova della pertinenza nella trattazione dei temi della riforma con presa di posizione netta sia a favore sia contraria, soprattutto nelle e-mail pertinenti "recuperate" che parlano della riforma a grandi linee.

Tabella 1 - I tre tipi di sentiment nei principali gruppi di e-mail

		Tipi di Sentiment				
		A - Aggettivi qualificativi		B - Parere		C - Propositività
Classificazione delle email		Positivo	Negativo	Contrario	Favorevole	Condizionale
Corpus complessivo	**%**	**76,4**	**42,6**	**12,9**	**17,7**	**15,5**
32.443 email	**val. ass.**	24.787	13.834	4.194	5.733	5.040
1) Riforma punto per punto	**%**	**88,8**	**56,3**	**23,4**	**26**	**23,2**
8.331 email	**val. ass.**	7.398	4.693	1.950	2.162	1.936
2) Riforma a grandi linee	**%**	**98,5**	**82,1**	**32,1**	**30,6**	**32,9**
925 email	**val. ass.**	869	724	283	270	290
3) Storie di vita	**%**	**83,1**	**50,1**	**13,9**	**15,5**	**22,3**
11.413 email	**val. ass.**	9.488	5.714	1.584	1.766	2.544
4) Petizioni pertinenti	**%**	**72,3**	**30,2**	**3,9**	**17,3**	**2,1**
8.216 email	**val. ass.**	5.943	2.484	318	1.421	174

5.4. Le dimensioni del Sentiment

Al fine di rendere più immediata l'interpretazione del dato, si sono calcolate le percentuali di e-mail che presentano ciascun criterio[10]. L'informazione riferita alla totalità delle e-mail nasconde fenomeni assai diversi, si è voluta quindi distinguere per i tipi già definiti, rispettivamente: 1) e-mail pertinenti che commentano la "riforma punto per punto"; 2) e-mail che commentano la "riforma a grandi linee"); 3) e-mail senza riferimenti diretti alla riforma e che ne parlano in chiave di narrazioni personali ("Storie di vita"); 4) e-mail pertinenti rispetto alla riforma ma inviate come semplici risposte a forme di adesione collettiva ("Petizioni"). In

[10] Si è tenuto conto del numero di citazioni per ogni criterio in ciascuna email, in una seconda misura che ha prodotto statistiche di valori medi per i punti più commentati della riforma, riportate nel report finale.

Tabella 1 si illustrano le valutazioni dei tre tipi di sentiment - A) qualificazioni positive vs negative, B) parere favorevole vs contrario, C) propositività.
In questo excursus emerge una tonalità sostanzialmente positiva del sentiment. Osservando le differenze fra i diversi tipi si evidenzia come nel corpus totale delle 32.443 mail, il sentiment di tipo A (misurato qui con la percentuale di e-mail contenenti almeno un aggettivo qualificativo) rivela una incidenza del 76,4% per il Positivo e del 42,6% per il Negativo.

Conclusioni

I modelli hanno risposto abbastanza bene quanto alla capacità di identificare i rispondenti e analizzare gli oggetti della consultazione. Inizialmente hanno permesso di depurare un primo grande blocco di mail arrivate, poco o per nulla pertinenti, in seguito la strategia che ha articolato differenti algoritmi ha consentito di captare la modulazione delle risposte arrivate. Infatti non tutti gli scriventi hanno risposto in modo pertinente all'invito lanciato dal Dipartimento della funzione pubblica, ossia contribuire con le proprie osservazioni ai 44 punti in discussione; fra coloro che li hanno trattati, ci sono soprattutto i funzionari con maggiori responsabilità. Inoltre coloro che, spesso con responsabilità minori o i cittadini comuni, pur "andando fuori tema", hanno partecipato attraverso il racconto delle loro esperienze, che abbiamo chiamato storie di vita. La consultazione sembra essere stata vissuta in modo positivo in particolare per i dirigenti presso cui c'era un interesse per aver avuto occasione di aver manifestato le loro priorità. E' sembrato che lo spirito dell'iniziativa è stata colta dai cittadini e dai funzionari della PA, e questo studio ha riscontrato una sostanziale positività nel materiale pervenuto, evento non scontato mentre ci si poteva aspettare un sentiment maggiormente negativo. Nonostante le lamentele sono rare quelle esclusivamente distruttive. La critica sterile e fine a sé stessa è pressoché assente. Chi ha risposto, e si è implicato, si è posto comunque come un interlocutore, anche quando non è stato al tema della consultazione.

Bibliografia

Battisti N. and Dolcetti F. (2012). Emozioni e testo: costruzione di risorse per il tagging automatico. In Dister A., Longrée D. and Purnelle G. editors, JADT 2012. 11es journées internationales d'analyse statistique des données textuelles. Université de Liège, pp. 95-107.

Bolasco S. (2013). *L'analisi automatica dei testi. Fare ricerca con il text mining*. Carocci editore - Studi Superiori.

Bolasco S. e Della Ratta Rinaldi F. (2004). Experiments on semantic categorisation of texts: analysis of positive and negative dimension. in G. Purnelle, C. Fairon, A. Dister, editors, JADT2004 Le Poids des mots. Actes des 7es journées internationales d'analyse statistique des données textuelle. Louvain-la- Neuve : Presses Universitaires de Louvain, vol (1): 202-210.

Bolasco S., Pavone P. (2010), Automatic Dictionary and Rule-Based Systems for Extracting Information from Text. In Data Analysis and Classification, a cura di Palumbo F. et al. Springer

Canzonetti A., Capo F.M., Ruocco V. (2006), Estrazione di informazione da una base documentale dell'AGCM con il software Taltac2: un esempio di integrazione fra strumenti di Text Mining e tecniche di data mining, in J. M. Viprey editor, JADT 2006 - Actes des 8es Journées Internationales d'Analyse Statistique des Données Textuelles, pp.245-252, Presses Universitaires de Franche-Comté, Besançon.

Carli R. and Paniccia R. M. (2002). *L'analisi emozionale del testo. Uno strumento psicologico per leggere testi e discorsi*. Milano: FrancoAngeli.

Osgood C.E., Suci, G. J., & Tannenbaum, P. H. (1957). *The measurement of meaning*. Urbana, University of Illinois Press.