

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

PROJECT WORK
NATURAL LANGUAGE PROCESSING

Linee guida per la classificazione della polarità delle frasi

Authors:

Francesco ANTICI
Luca BOLOGNINI
Matteo A. INAJETOVIC
Bogdan IVASIUK

Supervisors:

Paolo TORRONI
Federico RUGGERI
Andrea GALASSI



19 febbraio 2021

Indice

1	Introduzione	4
2	Perimetro	5
2.1	Scelta delle fonti	5
2.2	Split delle frasi	5
3	Studi e statistiche del dataset	7
3.1	Statistiche di classificazione	7
3.1.1	Kappa di Cohen	7
3.1.2	Kappa di Fleiss	8
3.1.3	Risultati ottenuti	9
4	Oggettività della frase	10
4.1	OGG 1	10
4.2	OGG 2	10
4.3	OGG 3	10
4.4	OGG 4	11
4.5	OGG 5	11
4.6	OGG 6	11
4.7	OGG 7	12
4.8	OGG 8	12
5	Soggettività della frase	13
5.1	SOG 1	13
5.2	SOG 2	13
5.3	SOG 3	13
5.4	SOG 4	14
5.5	SOG 5	14
5.6	SOG 6	15
5.7	SOG 7	15
6	Considerazioni sulle regole	16
6.1	Fact-checking	16
6.2	Frase che descrivono "emozioni"	16
6.3	L'utilizzo di parole superlative/non adatte al contesto	16
6.4	L'autore trae conclusioni senza la presenza di dati sufficienti	16
6.5	L'autore non trae conclusioni senza la presenza di dati	17

7	Regole per la classificazione degli articoli	17
7.1	Oggettività dell'articolo	17
7.2	Soggettività dell'articolo	18
8	Analisi del dataset	18
8.1	Statistiche delle frasi	18
8.2	Statistiche degli articoli	19
	Appendices	20

Sommario

In questo documento saranno riportate le linee guida stilate per la classificazione della polarità(Soggettiva/Oggettiva) di frasi in lingua italiana estratte da articoli di cronaca e decontestualizzate. Il fine di questo lavoro sarà la successiva creazione di un dataset di frasi classificate.

Tutto il materiale può essere consultato e scaricato direttamente da GitHub¹.

¹<https://github.com/francescoantici/ItalianNewspaperArticlePolarityDataset.git>

1 Introduzione

Nel campo del Natural Language Processing è fondamentale l'esistenza di materiale linguistico che possa essere utilizzato come base per le varie possibili applicazioni; ciò implica che per tutti i linguaggi naturali che si desiderino esplorare è necessario che questo tipo di dati siano disponibili. Questi, però, lo sono solamente in poche lingue, come ad esempio l'inglese; per le altre è necessario o creare nuovo materiale oppure tradurre da lingue più documentate.

Lo scopo di questo lavoro è di definire delle linee guida affidabili da seguire durante la creazione di questo materiale e generare un dataset italiano per studi relativi alla Subjectivity Analysis; perciò, sono state stilate delle regole per poter definire nella maniera meno arbitraria possibile se una frase sia soggettiva o oggettiva.

Come fonti dei dati sono stati scelti degli articoli provenienti da diverse testate giornalistiche che in seguito sono stati separati in frasi che conseguentemente sono state classificate manualmente seguendo le linee guida stilate. La classificazione infine si estende all'articolo nella sua interezza, associando un'ulteriore etichetta che ne rappresenta la polarità.

Data la difficoltà di separare ciò che gli annotatori stessi considerano soggettivo o meno, le regole sono state raffinate attraverso un lavoro di confronto iterativo collettivo sugli stessi articoli. Una volta raggiunta una buona concordanza interna sulla classificazione delle frasi, alle regole è stata data la loro forma finale e il dataset è stato creato assegnando a ciascun annotatore un egual numero di articoli, che sono stati analizzati individualmente.

2 Perimetro

2.1 Scelta delle fonti

Riconoscere la soggettività di una frase nella vita quotidiana risulta essere un compito generalmente semplice: approcciarsi a contesti in cui le opinioni sono solitamente apertamente dichiarate è pressoché immediato. Nell’ambito giornalistico, invece, prese di posizione forti e marcate tendono ad essere più rare. Con l’intento di coprire un ventaglio di argomenti trattati e registri linguistici, in modo tale che sia possibile analizzare possibili differenze tra essi, abbiamo preso in considerazione testate con caratteristiche differenti: testate di importanza nazionale e considerate generalmente come non schierate (rispetto a una fazione politica, per esempio) e testate locali o di opinionisti, come dei blog. Lo studio di conseguenza sarà meno dettagliato rispetto alle singole caratteristiche delle testate, ma coprirà uno spettro linguistico e uno stile autoriale più variegato, offrendo una visione più generale. Nell’appendice è possibile trovare i nomi delle testate e i link agli articoli presi in considerazione.

2.2 Split delle frasi

Come abbiamo detto all’inizio il dataset creato, al suo interno contiene sia la classificazione degli articoli che delle frasi. Quest’ultime sono state delineate seguendo delle regole precise, che vengono riportate in seguito.

1. La punteggiatura .?! delinea sempre la fine della frase;
2. l’apertura e la chiusura di una citazione con ”...”, « ... », -...-, rappresentano sempre la fine della frase;
3. ogni citazione non interrotta da punteggiatura viene considerata come frase;
4. l’utilizzo del ; per spezzare un periodo lungo rappresenta la fine della frase, al fine di evitare periodi troppo verbosi.

Al contrario:

1. l’utilizzo di -...-,”...”, « ... », in contesti non di citazione, ma di inciso non rappresenta la fine della frase;

2. l'utilizzo di " " per proverbi, appellativi, parole straniere, ecc.. non rappresenta la fine della frase.

La punteggiatura che interrompe la frase deve essere sempre eliminata dal testo prima di considerarlo parte del dataset. Il titolo, se esplicito e senza punteggiatura, viene considerato come frase a sé stante.

3 Studi e statistiche del dataset

In modo da ottenere una stima dell'efficacia delle regole concepite, è stato necessario valutare il risultato di ogni sessione di tagging collettiva attraverso una serie di studi statistici, solitamente molto comuni nella prassi di questa tipologia di lavoro. Questi studi consentono di rivedere in maniera critica le operazioni svolte manualmente dai componenti del gruppo di annotazione e valutarne la qualità collettiva.

3.1 Statistiche di classificazione

Le misure statistiche utilizzate per valutare la qualità delle regole nel guidare l'operazione di classificazione sono la Kappa di Cohen e la Kappa di Fleiss, che approfondiremo più nel dettaglio nei capitoli seguenti.

3.1.1 Kappa di Cohen

Il Kappa di Cohen è un coefficiente statistico che rappresenta il grado di accuratezza e affidabilità in una classificazione statistica. Questo indice di concordanza che tiene conto della probabilità di concordanza casuale: viene calcolato in base al rapporto tra l'accordo in eccesso rispetto alla probabilità di concordanza casuale e l'eccesso massimo ottenibile.²

Attraverso la matrice di confusione è possibile valutare questo parametro:

$$k = \frac{\Pr(a) - \Pr(e)}{1 - \Pr(e)} \quad (1)$$

dove

- $\Pr(a)$ - è data dalla somma della prima diagonale della matrice divisa per il totale dei giudizi e rappresenta la percentuale di giudizio, di fatto, concorde tra i giudici;
- $\Pr(e)$ - è il prodotto dei totali positivi sommato a quelli negativi, il tutto diviso per il quadrato del totale dei giudizi come rappresentato nella formula $(PP' + NN')/T^2$, e rappresenta la probabilità di accordo casualmente.

²Fonte wikipedia: https://it.wikipedia.org/wiki/Kappa_di_Cohen

Partendo dai valori assunti da k , che variano tra -1 e 1 , possiamo dedurre il "grado di concordanza" di due esaminatori. In seguito vediamo i possibili casi:

- $k \leq 0$ - non c'è concordanza;
- $0 < k < 0.40$ - concordanza scarsa;
- $0.41 < k < 0.60$ - concordanza discreta;
- $0.61 < k < 0.80$ - concordanza buona;
- $0.81 < k < 1$ - concordanza ottima;

3.1.2 Kappa di Fleiss

Il kappa di Fleiss è una misura statistica usata per valutare l'affidabilità dell'accordo tra un numero fisso di persone quando si assegnano valutazioni categoriali a un numero di elementi. Questa statistica entra in conflitto con altri kappa, come il kappa di Cohen ad esempio, il quale è attendibile solo quando si valuta l'accordo tra non più di due valutatori o l'affidabilità intra-valutatore (tra un valutatore e sé stesso). La misura calcola il grado di concordanza nella classificazione rispetto a quello che ci si aspetterebbe in uno scenario puramente casuale.³

$$k = \frac{\bar{P} - \bar{P}_e}{1 - \bar{P}_e} \quad (2)$$

dove

- \bar{P} - rappresenta l'accordo osservato;
- \bar{P}_e - rappresenta l'accordo atteso.

Come per Kappa di Cohen, sulla base dei valori di k possiamo valutare il "grado di concordanza" dell'intero gruppo.

- $k \leq 0$ - non c'è concordanza;
- $0 < k < 0.40$ - concordanza scarsa;

³Fonte wikipedia: https://en.wikipedia.org/wiki/Fleiss%27_kappa

- $0.41 < k < 0.60$ - concordanza discreta;
- $0.61 < k < 0.80$ - concordanza buona;
- $0.81 < k < 1$ - concordanza ottima;

3.1.3 Risultati ottenuti

In seguito possiamo vedere i risultati delle due *kappa* ottenuti grazie alle regole che vedremo nel dettaglio nei capitoli successivi. Le metriche sono state calcolate sia per le frasi che per gli articoli.

È importante sottolineare, che i risultati che vediamo, sono frutto di un lungo processo di affinamento delle regole. Infatti, i primi risultati ottenuti sono molto differenti.

Kappa di Cohen				
	Bogdan	Francesco	Luca	Matteo
Bogdan	-	0,02	0,37	0,27
Francesco	0,02	-	-0,07	0,06
Luca	0,37	-0,07	-	0,32
Matteo	0,27	0,06	0,32	-

Kappa di Cohen				
	Bogdan	Francesco	Luca	Matteo
Bogdan	-	0,53	0,54	0,52
Francesco	0,53	-	0,66	0,72
Luca	0,54	0,66	-	0,75
Matteo	0,52	0,72	0,75	-

Fleiss' Kappa	
	0,25

Fleiss' Kappa	
	0,62

Figura 1: A sinistra i risultati delle metriche sulle frasi dopo la stesura preliminare delle regole, a destra dopo la stesura definitiva.

Kappa di Cohen				
	Bogdan	Francesco	Luca	Matteo
Bogdan	-	0,40	0,58	0,40
Francesco	0,40	-	0,40	0,20
Luca	0,58	0,40	-	0,00
Matteo	0,40	0,20	0,00	-

Kappa di Cohen				
	Bogdan	Francesco	Luca	Matteo
Bogdan	-	0,47	0,39	0,47
Francesco	0,47	-	0,59	0,73
Luca	0,39	0,59	-	0,59
Matteo	0,47	0,73	0,59	-

Fleiss' Kappa	
	0,33

Fleiss' Kappa	
	0,53

Figura 2: A sinistra i risultati delle metriche sugli articoli dopo la stesura preliminare delle regole, a destra dopo la stesura definitiva.

4 Oggettività della frase

Di seguito riportiamo i criteri secondo i quali le frasi del dataset sono state considerate **Oggettive** (OGG).

4.1 OGG 1

Ogni fatto di cronaca/evento storico/dati/fenomeno citato è considerato come fattuale. In seguito alcuni esempi:

- *"La distanza tra Milano e Bologna è di 223 Kilometri"*
- *"28 dicembre 1977: i Nuovi Partigiani uccidono Angelo Pistolesi"*
- *"Dopo l'attrice, candidata a un Golden Globe per "Westworld" nel 2017 e prima ancora, nel 2003, per "Thirteen", altre quattro donne si sono fatte avanti per accusare il cantante di "Tainted Love" di simili violenze e abusi."*

4.2 OGG 2

Viene riportata un'opinione/condizione/emozione attribuibile esplicitamente a terzi. In seguito alcuni esempi:

- *"Senza scuola la giornata per Andrea è infinita"*
- *"I due inizialmente avevano concordato una parcella di 250.000 dollari per la difesa, ma Bowers non aveva incluso molte spese aggiuntive, tra cui nuovi legali e ricercatori impegnati nello studio del caso, che hanno fatto schizzare la parcella a 3 milioni di dollari, il che ha fatto infuriare Trump."*
- *"Il consumatore esprime tutta la sua amarezza sul web con un post."*

4.3 OGG 3

Lo scrittore con dati nella frase insufficienti non trae conclusioni. In seguito alcuni esempi:

- *"Non è ancora dato sapere"*

- *"Sembra anche che ci siano disponibilità di scorte a livello internazionale, ma è tutto da capire"*
- *"Nei prossimi giorni coinvolgeremo tutti i soggetti che potranno dare una mano per vaccinare 10 milioni di persone auspicabilmente entro l'inizio dell'estate, prima di giugno, ma sempre se ci danno i vaccini."*

4.4 OGG 4

Lo scrittore con dati sufficienti nella frase trae conclusioni ovvie o condivisibili alla luce dei fatti. In seguito alcuni esempi:

- *"Se la Cassazione rimandasse qualcuno degli imputati in appello e si rifacesse il processo, entro il 2021 cadrebbe prescritto anche il reato di disastro ferroviario"*
- *"I disoccupati complessivi sono 2.257.000 con un aumento di 34.000 unità su novembre e un calo di 222,000 su dicembre 2019."*
- *"Gli inattivi sono 13.759.000 e crescono di 42.000 unità su novembre di 482.000 unità su dicembre 2019."*

4.5 OGG 5

L'utilizzo di soprannomi di pubblico dominio è considerata oggettiva. In seguito alcuni esempi:

- *"Il cavaliere"* attribuito a Berlusconi
- *"Il duca bianco"* attribuito a David Bowie
- *"il Komandante"* riferito a Vasco Rossi

4.6 OGG 6

L'utilizzo di proverbi o modi di dire di pubblico dominio è considerata oggettiva. In seguito alcuni esempi:

- *"Ricordate il detto: La vendetta è un piatto che va servito freddo"*
- *"Solo il Papa è infallibile"*
- *"Chi va piano, va sano e va lontano"*

4.7 OGG 7

L'impossibilità di attribuire il tag di soggettività alla frase in questione. In seguito alcuni esempi:

- *"Di conseguenza hanno chiesto che il tribunale che si occupa del suo caso applichi la controversa condanna di tre anni e mezzo che aveva ricevuto nel 2014 per corruzione, e che era stata sospesa."*
- *"È toccato proprio a lei trovarlo ormai privo di vita."*
- *"Insomma, esiste ancora l'inviolabilità del domicilio."*

4.8 OGG 8

L'assenza di un'esplicita opinione personale. In seguito alcuni esempi:

- *"Tutti sono stati identificati e multati per aver violato le norme anti contagio per il contenimento del fenomeno epidemico."*
- *"Mamer è il capo degli addetti stampa della Commissione nonché il portavoce personale della presidente della Commissione, Ursula von der Leyen: quando parla ai giornalisti, lo fa a nome della Commissione."*
- *"Ha comportato una smentita secca del Quirinale quella secondo la quale Mattarella nei giorni scorsi avrebbe pre allertato Mario Draghi"*
- *"Domenica il dipartimento di polizia di Rochester ha diffuso il video della telecamera indossata da uno degli agenti (body camera) che riprende i fatti."*

5 Soggettività della frase

Ricordiamo che per valutare ogni frase la si deve considerare decontestualizzata rispetto alle altre frasi presenti nell'articolo di provenienza. La classificazione avviene sulla base dei criteri che vengono riportati in seguito.

5.1 SOG 1

La presenza di una qualsiasi opinione esplicitamente personale. In seguito alcuni esempi:

- *"Una mossa alla Angela Merkel, quella di Zaia"*
- *"C'è un margine anche in questo senso"*
- *"Hanno festeggiato il matrimonio come se non ci fosse il coronavirus."*

5.2 SOG 2

La presenza di un'espressione di sarcasmo/ironia. In seguito alcuni esempi:

- *"Beati loro che hanno solo questo problema."*,
- *"Ma non avevano detto di fare presto, che non si poteva perdere tempo?"*
- *"li abbiamo messi nelle aule dell'indirizzo scienze umane, ce li vede i ragazzi del classico a fare le versioni con il vocabolario su quei banchetti?"*

5.3 SOG 3

La presenza di esortazioni o auspici personali/di gruppo. In seguito alcuni esempi:

- *"Io mi auguro che Renzi lo quereli."*
- *"Auspicio tempi rapidissimi per gli esiti di quest'inchiesta, perché i valori democratici passano anche e, soprattutto, attraverso la terzietà e l'osservanza rigorosa delle norme da parte di coloro che ricoprono ruoli nella magistratura, cui spetta di salvaguardarli e tutelarli, così da*

non intaccare la fiducia che i cittadini nutrono nelle istituzioni e nella magistratura”

- *”A nome mio e dell’intera amministrazione comunale auguro buon lavoro alla dottoressa Matarrelli”*
- *”Come ripeto ormai da anni le soluzioni ci sono, sta a noi la scelta di sederci intorno ad un tavolo per riconvertire la qualità della nostra vita e creare un sistema che formi persone, non lavoratori.”*

5.4 SOG 4

La presenza di forme di discriminazione o declassamento. In seguito alcuni esempi:

- *”Voi siete il governo peggiore in assoluto.”*
- *”Sei un disco rotto.”*
- *”Ebrei nei forni vi bruceremo tutti”*
- *”Oggi Repubblica - solita fogna di regime - esce con un articolo dove millanta ipotetiche manovre di un nostro avvicinamento a Fratelli d’Italia.”*

5.5 SOG 5

L’utilizzo di parole superlative/non adatte ad un contesto oggettivo. In seguito alcuni esempi:

- *”decisione sulla crisi di governo entro le 15, un drammatico prolungamento”*
- *”Vaccini, pazzesco cambio in corsa?”*
- *”Un orrore firmato no-vax che si è scatenato dopo la sua ospitata di domenica sera a Non è l’Arena, il programma di Massimo Giletti su La7.”*
- *”E’ un disastro, un vero disastro”*

5.6 SOG 6

L'autore trae conclusioni senza la presenza di dati sufficienti. In seguito alcuni esempi:

- *"Insomma, per ora non è stato possibile acquistare autonomamente il siero, ma la via sembra comunque percorribile"*
- *"E così in queste concitate ore al capo ufficio stampa di Palazzo Chigi non rimane che tentarle tutte in quello che sa meglio fare: indirizzare i consensi"*
- *"Il punto più critico è a Siete, per il cui controllo sono stati bloccati i combattimenti ma proprio da quella località i mercenari russi alleati di Haftar hanno costruito una serie di fortificazioni che sembrano più una preparazione per la ripresa del conflitto che una proiezione futura della pace."*

5.7 SOG 7

L'espressione di una qualsiasi emozione / stato d'animo. In seguito alcuni esempi:

- *"Ansia, dolore, rabbia, si mescola tutto"*
- *"È come le altre volte, non dormo"*
- *"Gli occhi tristi di Billy mi danno un fortissimo stimolo per guarire."*
- *"Un'esperienza che ho cominciato come un sogno ed è finito in un incubo che mi rincorrerà per tanto tempo"*

6 Considerazioni sulle regole

6.1 Fact-checking

Ogni fatto riportato all'interno dell'articolo è stato considerato attendibile. Questo significa che non è stata effettuata nessuna analisi della veridicità delle notizie.

6.2 Frasi che descrivono "emozioni"

L'espressione di emozioni all'interno di una frase è stato motivo di confusione durante lo sviluppo delle regole: all'interno di frasi si può trovare riferimento alle emozioni provate, il che è un argomento estremamente soggettivo. Considerando che una frase che descrive quello che un soggetto terzo prova riporta la situazione che affronta, anche alla luce dell'assunta veridicità dei fatti riportati, la frase è considerata Oggettiva. Se le emozioni sono espresse direttamente dal soggetto, la frase è considerata Soggettiva poiché è il diretto interessato che dà voce a ciò che prova.

6.3 L'utilizzo di parole superlative/non adatte al contesto

L'utilizzo di parole "non adeguate" al contesto è un tema abbastanza spinoso, ma abbiamo voluto includerlo nelle nostre linee guida: una frase oggettiva deve anche avere un linguaggio conforme al contesto. Nel caso in cui vengano usate parole superlative, esagerate o che celino un'idea dello scrittore, tutta la frase viene considerata soggettiva al fine di dare una rilevanza assoluta anche alle parole usate e non solo alla forma.

6.4 L'autore trae conclusioni senza la presenza di dati sufficienti

Il caso di soggettività 6 è stato uno dei più dibattuti nella stesura delle regole. Pur rispettando sempre l'assunzione iniziale secondo cui i fatti riportati negli articoli vengono assunti come veri, nelle varie fasi di classificazione abbiamo affrontato frasi caratterizzate dalla presenza di relazioni di causa effetto non esplicitate tra dati esposti e conclusioni, associabili dunque ad un intervento "soggettivo" dell'autore.

6.5 L'autore non trae conclusioni senza la presenza di dati

La regola di oggettività 3 può essere considerata l'altra faccia della medaglia della regola trattata nel paragrafo precedente: in questo caso l'autore però non trae conclusioni partendo da una base insufficiente di dati, ma offre invece un possibile scenario evolutosi da essi. La decisione di considerare questo tipo di frasi come oggettive è dovuta al fatto che l'autore costruisce un'ipotesi, ma non pone mai l'accento sulla sua veridicità, quindi non trasmette la sua opinione a riguardo, lasciando la discussione aperta.

7 Regole per la classificazione degli articoli

La stesura di regole generali e oggettive quando si viene a considerare un articolo nella sua interezza si è rivelata più complicata di quella per le singole frasi: bisogna infatti considerare il messaggio complessivo dell'articolo e risulta più difficile delineare delle linee guida che risultino sufficienti a inquadrarlo in maniera oggettiva. Di seguito sono riportate le regole utilizzate per la classificazione degli articoli.

7.1 Oggettività dell'articolo

Un articolo è considerato oggettivo quando, leggendo contestualmente e complessivamente le frasi che lo compongono, la soggettività dell'autore non viene colta o percepita esplicitamente tramite prese di posizione o affermazioni personali significative.

Importante sottolineare che articoli caratterizzati da una presenza consistente di citazioni di terzi hanno frequentemente un numero elevato di frasi taggate come soggettive. Nonostante ciò, nel caso in cui l'autore non fa altro che riportare affermazioni altrui senza lasciar trasparire opinioni e percezioni personali, l'articolo viene considerato oggettivo. La decisione presa è in linea con l'idea generale dell'applicare il tagging di oggettività/soggettività in relazione all'autore stesso e non a terzi, coerentemente con la regola riportata nel paragrafo 4.2.

7.2 Soggettività dell'articolo

Un articolo è considerato soggettivo se leggendolo e contestualizzando tutte le sue frasi si riesce a cogliere la posizione dello scrittore. Questa può essere espressa esplicitamente, tramite commenti e valutazioni personali, ma anche tramite una disparità di trattamento delle parti in causa nella narrazione di un fatto (es: citazioni molteplici di solo una delle due parti in causa). Possiamo concludere che la classificazione soggettiva avviene nelle casistiche opposte a quelle descritte nel paragrafo 7.1, ossia se caratterizzato da prese di posizione marcate, considerazioni emotive, esplicite manifestazioni di supporto o di critica rispetto ai temi, ideologie o personaggi di dominio pubblico, da parte dell'autore dell'articolo in questione.

8 Analisi del dataset

Al fine di valutare ed osservare il dataset ottenuto abbiamo effettuato una serie di analisi statistiche. In alcuni casi, le statistiche ottenute possono essere sfruttate per massimizzare i risultati di alcuni esperimenti, come ad esempio la ricerca dei parametri ottimali.

8.1 Statistiche delle frasi

In seguito riportiamo una serie di statistiche delle frasi classificate. Il totale delle frasi è 1841. Nella 11 possiamo vedere la quantità di frasi in termini percentuali rispetto al totale relativo di ciascun annotatore. E' possibile notare come il livello di frasi soggettive oscilla in un range tra 20% e 40%, facendo leva sul fatto che ogni annotatore doveva prendere in considerazione testate più o meno note e più o meno autorevoli. Analizzando le singole testate giornalistiche incluse nel nostro esperimento abbiamo ottenuto un altro risultato interessante: come si può vedere dalla 5 e 4, abbiamo in termini percentuali una quantità di frasi rispetto al totale complessivo per ciascuna testata giornalistica. Questo dato è naturalmente influenzato dal totale di frasi analizzate per ciascuna testata. Infine abbiamo deciso anche di valutare se ci fosse o no prevalenza di una regola applicata per classificare le frasi soggettive. Possiamo notare dalla 10 che per la maggior parte dei casi, le frasi che sono state valutate soggettive contenevano un'opinione esplicita al loro interno.

8.2 Statistiche degli articoli

Data la presenza delle regole sia per la classificazione delle frasi che degli articoli, abbiamo voluto indagare anche sui dati relativi ai singoli articoli per capire la presenza di qualche bias oppure correlazione tra la testa giornalistica e soggettività degli articoli.

Nella 8 vediamo quanti sono gli articoli soggettivi ed oggettivi per ciascuna testata giornalistica. Inoltre abbiamo analizzato se ci fosse qualche relazione tra la lunghezza dell'articolo e la sua classificazione, per fare questo abbiamo suddiviso per lunghezza gli articoli in 4 classi. Come possiamo vedere dalla figura 9, con l'aumentare della lunghezza degli articoli, aumenta la presenza di soggettività. Gli articoli brevi risultano essere raramente soggettivi.

Appendices

A First section

Fonte
Il Fatto Quotidiano
Libero
Contropiano
Il Giornale
Espresso
Repubblica
Corriere
Beppe Grillo
Esse notizie
il post
Il Foglio
Avanti
ANSA
Avvenire
FascInAzione
Unità

B Second section

B.1 Statistiche delle frasi

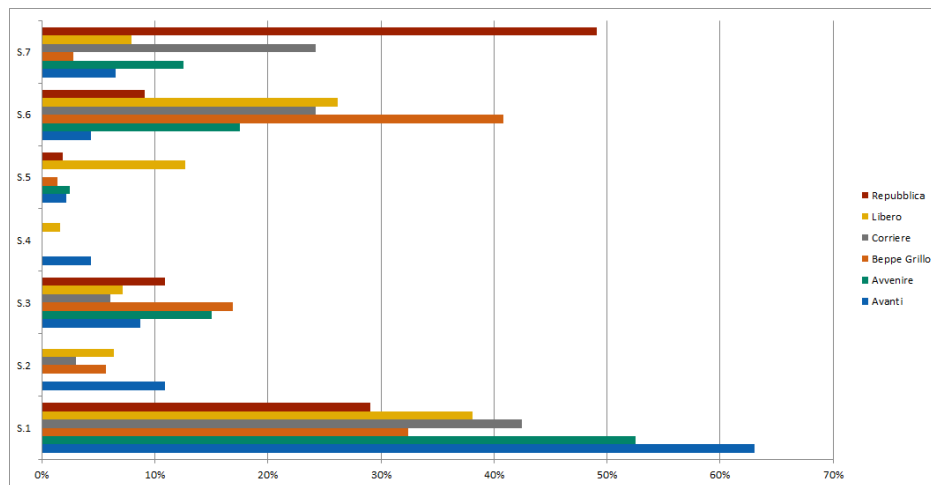


Figura 3: Regole applicate in termini percentuali negli articoli soggettivi

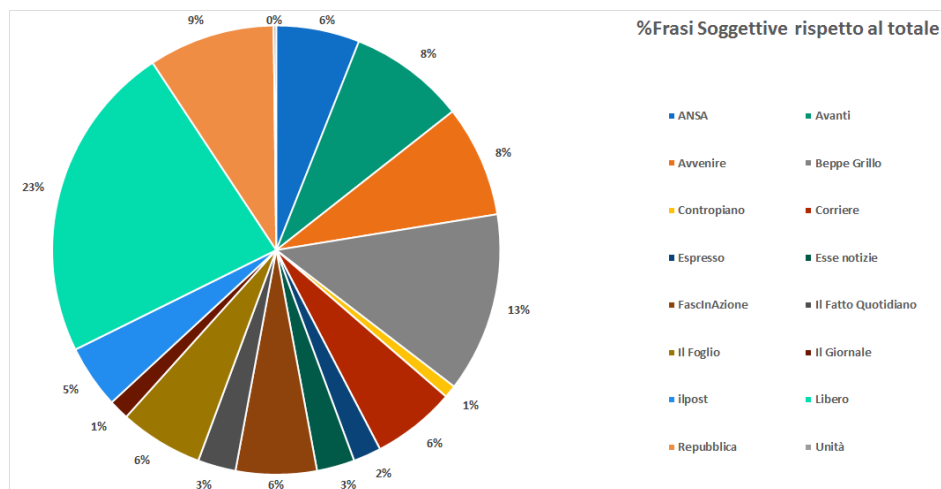


Figura 4: Percentuali delle frasi soggettive all'interno degli articoli

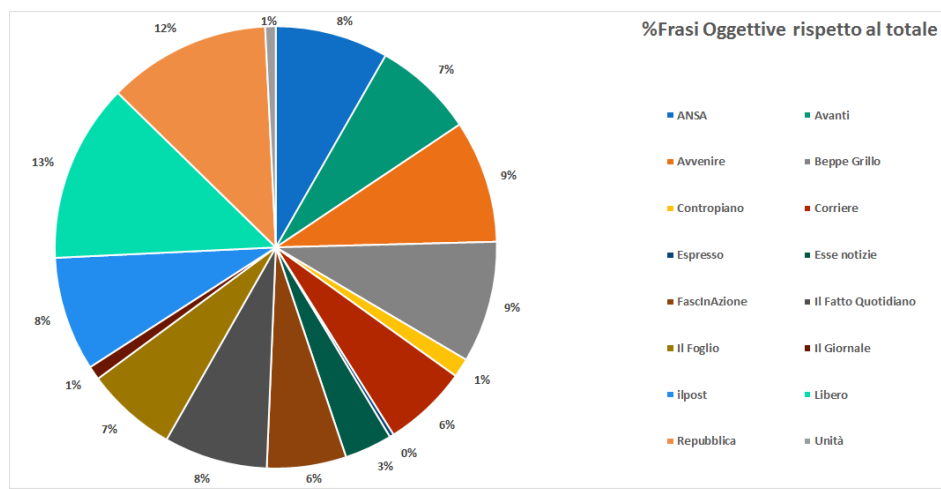


Figura 5: Percentuali delle frasi oggettive all'interno degli articoli

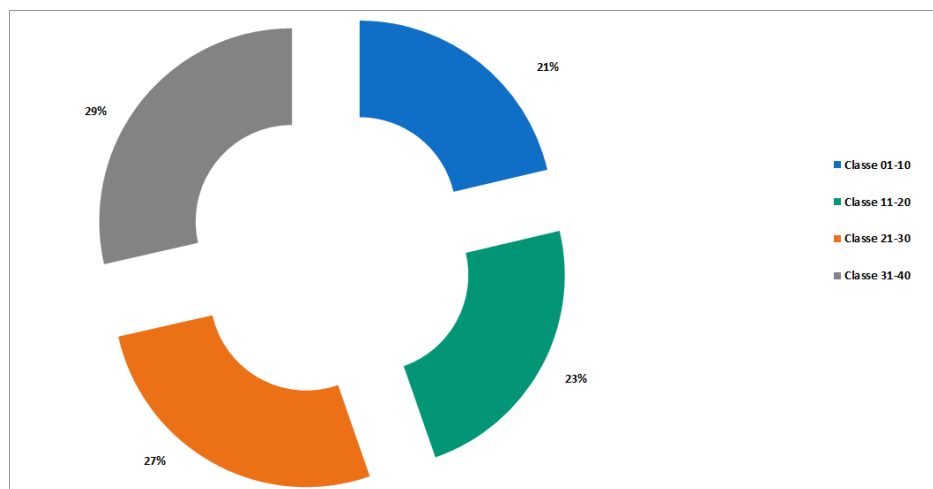


Figura 6: Percentuali delle frasi all'interno degli articoli soggettivi

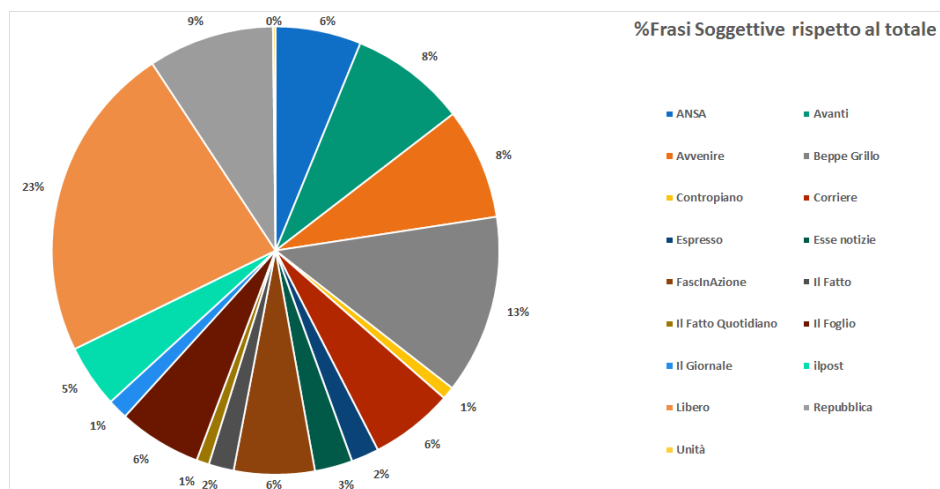


Figura 7: Percentuali delle frasi all'interno dei singole testate

B.2 Statistiche degli articoli

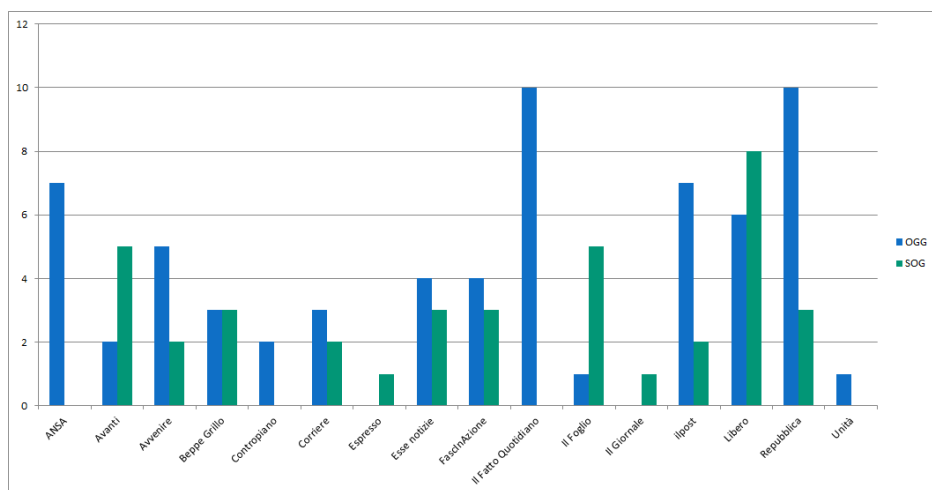


Figura 8: Classificazione degli articoli

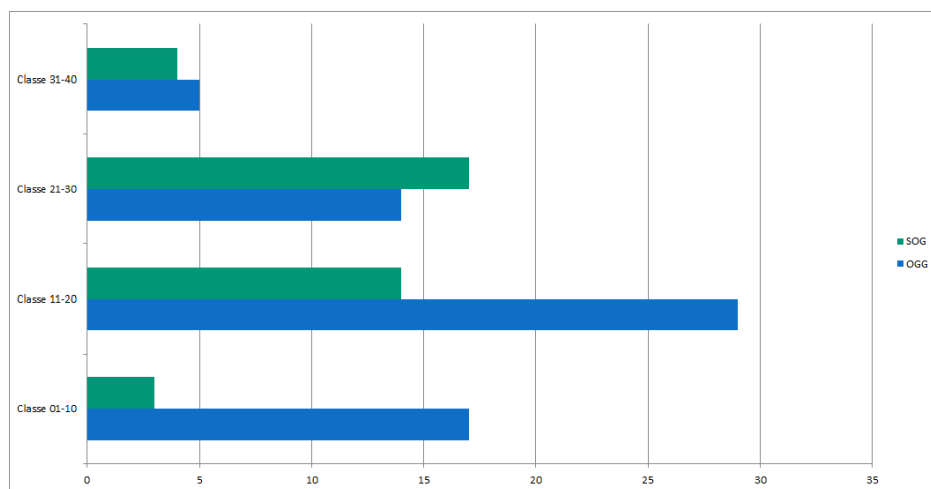


Figura 9: Classificazione degli articoli in base alle classi assegnate

B.3 Statistiche delle persone

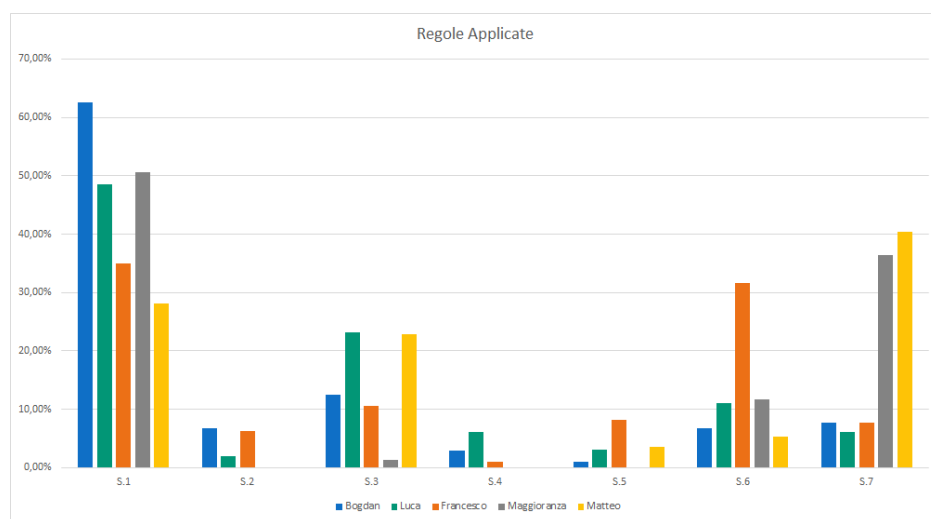


Figura 10: Regole applicate in termini percentuali

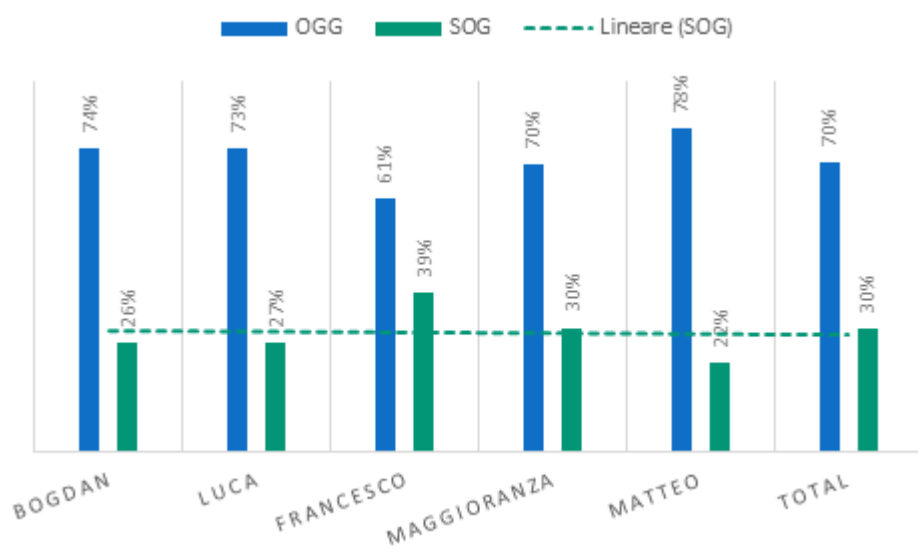


Figura 11: Classificazione delle frasi dei singoli annotatori