2. Parole e numeri: la linguistica dei corpora come metodo per individuare le bufale

Elia Silvestro

Le bufale si possono distinguere dalle notizie vere. Fin qui nulla di strano: in certi casi le sparano talmente grosse che è improbabile cadere nella trappola. La vera notizia, però, è un'altra: se si hanno gli strumenti adatti, per smascherarle non serve nemmeno leggerle. L'affermazione, che può suonare inverosimile, è il risultato delle ricerche sulle notizie false che illustreremo in questo capitolo. La domanda da cui siamo partiti era: possiamo distinguere le notizie false da quelle vere per *come* sono scritte più che per *che cosa* sostengono? In altre parole, ci sono delle caratteristiche linguistiche che permettono di individuare una bufala senza perdere tutto il tempo necessario a verificarne i contenuti?

Come vedremo, la risposta è affermativa, e questa è la conclusione a cui siamo giunti grazie agli strumenti di analisi offerti dalla linguistica dei corpora, una branca della linguistica che potrebbe rivelarsi di fondamentale importanza in un'epoca, come quella odierna, caratterizzata da un sovraccarico di informazioni. Tramite strumenti informatici diversi siamo riusciti ad analizzare grandi quantità di articoli veri e falsi in tempi ridotti, concentrandoci sull'individuazione delle loro caratteristiche formali. Inoltre, abbiamo cercato di capire

se ci fosse qualche differenza tra le bufale ingannevoli (quelle che si spacciano per notizie veritiere) e gli articoli parodistici (dichiaratamente ironici, scritti per lettori consapevoli in cerca d'intrattenimento).

2.1. Piccolo vademecum per 'sbufalare' una notizia falsa

All'atto pratico, va ammesso, non è vero fino in fondo che non serve leggere una notizia sospetta per stabilire che è una bufala: sarebbe possibile solo progettando un sistema informatico in grado di analizzare gli articoli al posto nostro. Piuttosto, ciò che possiamo fare come lettori è allenarci a riconoscerne rapidamente le caratteristiche linguistiche e, di conseguenza, a farci sospettare di un articolo anche quando il suo contenuto ci sembra plausibile. A questo scopo, per prima cosa abbiamo redatto una breve guida per individuare una bufala partendo dai suoi tratti linguistici più evidenti. Più avanti nel capitolo forniremo una spiegazione teorica più dettagliata del percorso di ricerca che ha permesso di arrivare a queste indicazioni.

Qualunque sia l'argomento trattato, lo stile delle bufale le rende riconoscibili per certi tratti, che invece risultano inconsueti nei normali articoli di giornale. È dunque fondamentale individuare i principali campanelli d'allarme che ci fanno insospettire di fronte a una notizia di dubbia origine.

2.1.1. Manteniamo le distanze

Ciò che ci aspettiamo da un articolo di giornale è che si concentri su un certo argomento, e solo secondariamente sui lettori. Gli appelli diretti sono riservati ad alcune tipologie specifiche di articoli, come gli editoriali, cioè gli articoli in cui il direttore del giornale, un giornalista famoso o un esperto esprime le proprie opinioni su un dato evento o argomento. Questa, però, non è una tipologia di articoli in cui si manifestano le bufale, che quasi sempre sono firmate con pseudonimi e riportano notizie precise o illustrano una novità. L'uso di verbi alla seconda persona, tra cui gli imperativi (*guardate!*, *leggete!*, *riflettete!* ecc.), in articoli che non rappresentano l'opinione dell'autore, indica l'inosservanza di una precisa regola giornalistica e può essere il segno che la notizia che stiamo leggendo è falsa.

L'uso della seconda persona è solo uno dei vari modi che gli autori di bufale hanno a disposizione per attirare la nostra attenzione o suscitare il nostro coinvolgimento emotivo. Un altro è l'uso di iperboli: parole che sovraccaricano la descrizione dei fatti, esagerandone le circostanze. È davvero necessario descrivere un crimine come *allucinante* o una scelta di politica economica come uno *shock*? Le scoperte sul cancro di misteriosi ricercatori devono sempre essere *incredibili*? Non per forza, anche perché parole simili andrebbero oltre l'intento primariamente informativo di un giornale autorevole. Le notizie false, invece, spingono il lettore a reagire (con rabbia, gioia, stupore) in modo da invogliarlo a condividere i suoi sentimenti con i contatti che ha sui social network o nella vita reale.

2.1.2. Argomenti forti, contenuti deboli

Uno dei segnali più chiari del fatto che ci troviamo di fronte a una notizia inventata è la povertà di dettagli. È esperienza comune la difficoltà di inventarsi una frottola che regga anche nei suoi particolari: più ne aggiungiamo, più rischiamo di far crollare il castello in aria che abbiamo costruito. Gli autori di bufale, invece, hanno tutto l'interesse a creare canovacci narrativi riutilizzabili all'infinito, come accade per le fiabe: la storia è fondamentalmente sempre quella, cambiano solo un paio di dettagli. In una bufala a tema politico, per esempio, troveremo appena un paio di nomi delle figure più di spicco e qualche vago riferimento a una legge; poche tracce di date, indicazioni geografiche o in generale di riferimenti all'attualità stretta. Gli argomenti sono scelti tra gli *evergreen*, quelli che interessano qualsiasi lettore a prescindere da età, professione o provenienza (per esempio tasse ingiustificate, privilegi della classe dirigente e così via), in modo da risultare sempre attuali e non avere bisogno di un legame preciso col dibattito politico o con la cronaca del periodo.

2.1.3. Errori e orrori

Se notare ciò che manca (come illustrato sopra) non è semplice, possiamo comunque concentrarci su alcuni segnali ben visibili. Le bufale hanno uno stile piuttosto riconoscibile: spesso contengono delle caratteristiche linguistiche che potremmo incontrare

in una conversazione tra amici, ma che non ci aspetteremmo di trovare in un articolo pubblicato da un giornale.

Un segnale chiaro che l'articolo che stiamo leggendo potrebbe essere una bufala è la presenza di errori ortografici e punteggiatura insolita. È bene aguzzare la vista in cerca di doppie mancanti o sovrabbondanti, accenti dimenticati o sostituiti da apostrofi, ma anche di punteggiatura imprecisa (per esempio due puntini .. invece di tre ...) o sovrabbondante (due o più punti esclamativi *!!* o interrogativi *??* di seguito non vengono mai usati in articoli di giornale scritti da professionisti).

2.1.4. Parla come mangi (prima parte)

Se invece guardiamo alla scelta delle parole, la presenza di lessico colloquiale, o a sfondo sessuale, e del turpiloquio ci possono aiutare a scovare le bufale. Sono perlopiù gli articoli parodistici, cioè quelli dichiaratamente ironici, ad abbondare di queste parole che, per ovvi motivi, non troverebbero spazio su un giornale autorevole. Tuttavia, anche nelle bufale ingannevoli (quelle che non dichiarano esplicitamente di essere notizie false), emergono talvolta scelte linguistiche che non rispettano il "politicamente corretto" (per esempio *barbona* invece di *senzatetto*, *zingari* invece di *etnia rom*, *negro* al posto di un'indicazione geografica o etnica).

Troviamo anche altre parole che, pur non essendo inopportune come gli insulti o le offese, sarebbe strano trovare in un articolo veritiero. Si tratta del lessico colloquiale e informale, quello che usiamo normalmente in una conversazione tra amici ma che non inseriremmo in una lettera formale (per esempio *sfigati* invece di *sfortunati*). Un'altra categoria di parole colloquiali, meno evidente ma altrettanto indicativa, è quella dei verbi 'procomplementari': si tratta di un termine specialistico che indica quei verbi che all'infinito non terminano per *-are*, *-ere*, *-ire*, ma contengono dei pronomi. Verbi come *farcela*, *smetterla*, *fregarsene*, *tirarsela* sono tipici del parlato e non vengono normalmente utilizzati dagli autori di quotidiani autorevoli. Lo stesso discorso vale per le espressioni regionali o dialettali: se in un articolo troviamo parole come *ciucco* o *umarell*, è probabile che non si tratti di una notizia a cui prestare fiducia.

È chiaro, insomma, che un articolo che sembra troppo informale

per trovarsi su un giornale deve insospettirci (ovviamente questo vale anche per i testi pubblicati online). In pochi e limitati casi, tuttavia, vale l'opposto: anche alcune parole molto formali e cadute in disuso possono segnalarci che un articolo è una bufala. Pronomi come *egli*, *ella*, *esso*, *essa*, *essi* o *ciò*, *tale*, *costoro* sono sì formali, ma soprattutto desueti, tanto che ormai li incontriamo solo in qualche vecchio testo di grammatica o nei temi scolastici; sicuramente non in un articolo di giornale, che usa uno stile più moderno. Chi li usa, magari insieme ad altre scelte lessicali ricercate, probabilmente vuole rafforzare la nostra illusione di leggere un articolo affidabile con paroloni indecifrabili, quando invece sono i quotidiani stessi a evitarli per rendere i loro articoli più leggibili. Vediamo un esempio di questo fenomeno tratto dal corpus di bufale analizzato:

> 1) I cittadini italiani [...] diverranno, sotto riserva di quanto dispone il paragrafo seguente, cittadini godenti di pieni diritti civili e politici dello Stato al quale il territorio viene ceduto, secondo le leggi che a tale fine dovranno essere emanate dallo Stato [...]. Essi perderanno la loro cittadinanza italiana al momento in cui diverranno cittadini dello Stato subentrante.

2.2. Provare per credere: l'analisi quantitativa degli articoli

Se sapere quali segnali linguistici ci permettono di riconoscere una bufala è fondamentale, si tratta comunque solo della punta dell'iceberg della ricerca. Per capire meglio come si può arrivare a individuare questi segnali e sviluppare ulteriormente il nostro senso critico di lettori, può essere utile soffermarci su metodi e dati, cioè su *come* e *perché* i numeri ci aiutano a indagare la lingua in modo efficace. Ed è quello che faremo, appunto, nelle pagine che seguono.

2.2.1. I numeri e le bufale

Per verificare se fosse possibile distinguere le bufale dagli articoli di giornale veritieri partendo dalle loro caratteristiche formali, è stato utile un approccio *quantitativo* allo studio della lingua, che in ambito accademico viene definito "linguistica dei corpora". Si tratta, in sostanza, dell'uso di mezzi informatici per indagare

raccolte di testi sufficientemente grandi da permettere di notare delle tendenze linguistiche di fondo che potrebbero sfuggire al singolo lettore. In altre parole, il computer ci aiuta a capire se esistono differenze numericamente significative che distinguono due tipi di testi (nel nostro caso, le bufale e gli articoli veritieri). Per esempio, quale dei due tipi presenta un maggior uso di punti esclamativi, una lunghezza delle frasi leggermente minore rispetto alla media o un certo tempo verbale usato molto spesso? Dopo aver individuato queste caratteristiche, è fondamentale cercare di capire anche le ragioni che le determinano: perché troviamo certi fenomeni e non altri?

A questo punto occorre però soffermarci sull'utilità dell'approccio basato sulla linguistica dei corpora. In effetti, tabelle, formule e percentuali sono di norma concetti difficili da associare alla lingua. Perché affidarsi a uno strumento così arido rispetto a metodi apparentemente più familiari, come la lettura e l'analisi approfondita dei singoli articoli, in altre parole all'analisi *qualitativa*? La risposta sta nelle potenzialità dell'analisi quantitativa. Se è vero che le intuizioni (qualitative) che abbiamo leggendo anche un solo testo possono essere un buon punto di partenza, estendere questo metodo a raccolte più ampie è impossibile all'atto pratico per un semplice motivo: ci vorrebbe troppo tempo per leggere tutti i testi. Nel caso delle notizie false, esistono piattaforme online di *fact checking* (cioè di controllo della bontà dei contenuti delle notizie), come per esempio il portale *bufale.net* (di cui parleremo più avanti), i cui autori si dedicano volontariamente alla verifica dei contenuti di articoli ritenuti sospetti. Tuttavia, è evidente che pensare di poter rincorrere il flusso di notizie che travolge il web ogni giorno sarebbe illusorio. Per esempio, l'istituzione della "Giornata del *fact checking*" per il 2 aprile, l'indomani del giorno tipicamente dedicato ai "pesci d'aprile" (a volte spiritosi, talora insinuanti), è indubbiamente un'iniziativa lodevole come tentativo di sensibilizzare il pubblico, ma di efficacia limitata.

La trasformazione delle caratteristiche linguistiche delle bufale in valori numerici, invece, ci permette di lasciare il lavoro pesante all'analisi automatica. Una volta individuati i tratti linguistici che differenziano le bufale dai veri articoli di giornale, è sufficiente programmare degli strumenti informatici perché li ricerchino in nuovi testi e ci forniscano un'indicazione di quanto questi ultimi si avvicinino al modello di una notizia falsa.

2.2.2. I corpora

Per questa ricerca abbiamo innanzitutto individuato delle fonti disponibili in forma digitale che corrispondessero il più possibile agli standard di articoli di giornale autorevoli, notizie false ingannevoli e articoli parodistici. La scelta è ricaduta su tre fonti, dalle quali abbiamo raccolto dei corpora, cioè delle raccolte molto voluminose, di articoli:

- 100.000 parole di notizie veritiere dall'archivio di Repubblica.it (*ricerca.repubblica.it*), che offre tutti gli articoli del quotidiano in forma digitale dal 1984 a oggi;
- 50.000 parole di notizie false ingannevoli dal sito Bufale.net (*www.bufale.net*), una nota piattaforma italiana di *fact checking*: si tratta di un sito in cui contenuti ritenuti inaffidabili vengono verificati e viene stabilita la verità dei fatti. Questa è stata la parte più difficile della raccolta: in molti casi la forma dei testi è troppo disomogenea per poterla confrontare con un articolo di giornale (per esempio, manca il titolo se si tratta di catene di Sant'Antonio diffuse tramite applicazioni di messaggistica). Anche i problemi tecnici hanno avuto il loro ruolo: i portali che pubblicano notizie false hanno spesso vita breve, innanzitutto per problemi legati ai loro discutibili contenuti, ma anche perché gli articoli vengono spesso riutilizzati a distanza di poco tempo, dopo aver modificato i dettagli di un canovaccio che è sempre lo stesso.
- 50.000 parole di notizie false parodistiche dal sito Lercio.it (*www.lercio.it*), un famoso portale di notizie inventate a scopo d'intrattenimento, così popolare sui social network italiani da poter essere considerato l'esempio più celebre di questo genere pseudo-giornalistico.

Affinché i risultati delle analisi fossero affidabili, era importante creare dei corpora omogenei, ovvero stabilire una possibilità di confronto a livello di temi, arco temporale, forma e lunghezza degli articoli. Abbiamo quindi stabilito alcune regole per la raccolta.

Abbiamo incluso in percentuali simili articoli di diversi argomenti: politica/mondo, cronaca, sport, cultura e spettacolo, medicina/scienza/internet. Ovviamente, l'argomento di un articolo incide sul lessico utilizzato, e l'obiettivo era di avere un lessico il più simile possibile tra i diversi corpora (d'ora in avanti indicati per comodità come *Repubblica, Bufale* e *Lercio*). La composizione percentuale per argomenti che ne risulta è riportata nella tabella che segue.

	Repubblica	Bufale	Lercio
Politica/ Mondo	32,61%	34,23%	31,09%
Cronaca	40,87%	38,26%	37,31%
Sport	4,35%	1,34%	4,15%
Cultura/ Spettacolo	13,04%	15,44%	15,03%
Medicina/ Scienza/ Internet	9,13%	10,74%	12,44%

Tabella 1 – Distribuzione dei contenuti nei subcorpora (%)

Sempre per fare sì che gli argomenti esposti fossero il più possibile compatibili, abbiamo stabilito un periodo entro il quale doveva rientrare la pubblicazione degli articoli (dal febbraio 2014 all'ottobre 2015), ipotizzando che in questo arco di tempo i fatti di cronaca coperti dalle notizie fossero circa gli stessi.

Abbiamo eliminato tutte le parti degli articoli, come titolo, sottotitolo, occhiello, che non fanno parte del corpo del testo. Queste componenti, che vengono definite dagli studiosi *paratesto* (ciò che sta attorno al testo), hanno caratteristiche linguistiche proprie, che rischiavano di viziare i risultati dell'analisi (per esempio, nei giornali è la redazione e non l'articolista che decide il titolo di un pezzo). Inoltre, durante la ricerca abbiamo notato che gli articoli di *Bufale* e di *Lercio* si limitavano in massima parte al titolo e null'altro, al contrario degli articoli di *Repubblica*, ricchi di altri elementi paratestuali (occhiello, sottotitolo, sommario). Su titoli e sottotitoli torneremo poi separatamente più avanti, con un'analisi prettamente qualitativa.

I nostri corpora comprendono solo articoli di cronaca di lunghezza media, e non articoli di analisi e commento (editoriali, rubriche, elzeviri e così via), solitamente molto più lunghi degli altri e caratterizzati da uno stile più personale. Un'eccezione è rappresentata

dagli articoli di divulgazione medico-scientifica, che in *Bufale* erano sostanzialmente più lunghi. La tabella sotto riporta la lunghezza media in parole dei testi raccolti.

	Repubblica	Bufale	Lercio
Medicina/ Scienza/Internet	433,67	685,96	349,33
Cronaca	438,05	273	375,38
Media generale	437,65	336,74	371,90

Tabella 2 - Lunghezza media in parole dei testi raccolti

Ancor prima di analizzarne le parole, lo scarto nella lunghezza media degli articoli veritieri e falsi mostra già una certa differenza, particolarmente evidente se ci concentriamo sugli articoli di cronaca: le bufale sono notevolmente più brevi degli articoli di *Repubblica*, con *Lercio* in posizione intermedia. Un'ipotesi, che ci viene suggerita anche dalle analisi sul lessico che vedremo più avanti, è che le notizie false, in quanto inventate, siano difficili da arricchire di dettagli in quantità sufficiente da renderle paragonabili in lunghezza a degli articoli veritieri. La loro natura di testi redatti per il web, inoltre, potrebbe spingere i loro autori a scrivere articoli più snelli, adatti a un consumo veloce, canalizzato dai clic ottenuti attraverso le condivisioni sui social network.

2.2.3. Gli strumenti

Per ottenere dei dati da analizzare abbiamo "dissezionato" le raccolte di articoli per mezzo di diversi strumenti informatici. Abbiamo proceduto a tre confronti distinti: uno tra le 100.000 parole di articoli di *Repubblica* (lo chiameremo *Repubblica totale* da qui in poi) e le 100.000 parole di notizie false, e altri due tra 50.000 parole selezionate dagli articoli di *Repubblica* e ciascuno dei due corpora di notizie false, quelle ingannevoli (tratte da *Bufale.net*) e quelle parodistiche (tratte da *Lercio.it*).

Abbiamo iniziato analizzando ciascuna coppia di corpora con *TalTac*^{2}, un software sviluppato dalla Sapienza - Università di Roma. *TalTac*^{2} analizza da cima a fondo tutte le forme lessicali (le singole parole, come "tavolo" o "che") e i segmenti (sequenze di forme che

si presentano in un determinato ordine più volte, come "durante la notte") contenuti in un corpus e li riordina in una tabella indicando quelli più frequenti e quelli più rari. Permette, inoltre, di calcolare alcuni valori utili a capire quante parole diverse gli autori usino per scrivere i testi analizzati: si tratta di ciò che in linguistica viene definita "ricchezza lessicale". Una volta fatti i calcoli, *TaITac*[2] permette di incrociare i dati, rivelando quale e quanto lessico due corpora hanno in comune e quale tra i corpora usa un lessico più vario.

Grazie a un altro software, *TreeTagger*, sviluppato dall'Università di Stoccarda, abbiamo ottenuto i dati riguardanti la "densità lessicale" di ciascuno dei corpora, cioè la frequenza di parole che hanno di per sé un significato (che gli esperti chiamano "parole piene": aggettivi, avverbi, nomi e verbi) o che invece servono principalmente a costruire la struttura sintattica della frase (le "parole vuote": articoli, congiunzioni, interiezioni, preposizioni e pronomi). *TreeTagger* attribuisce automaticamente a una di queste categorie ciascuna delle parole e ci ha permesso di calcolare la percentuale di ciascuna categoria in ognuno dei corpora. I dati sulla densità lessicale sono un'ottima misura dello stile dei testi: ci permettono, in particolare, di capire quanto un testo sia simile alla lingua parlata (di norma più ricca di parole vuote) oppure alla lingua scritta (caratterizzata di solito da una maggior presenza di parole piene).

Tramite *Corrige!*, una piattaforma online di analisi automatica dei testi, abbiamo ricavato un resoconto ortografico (una lista di errori di battitura, parole imbarazzanti, errori di punteggiatura ecc.) e un'analisi della facilità di lettura del testo (la "leggibilità"), basata sia sulla lunghezza delle parole e delle frasi, sia sulla complessità del lessico presente nei corpora. Questi dati hanno permesso di confermare quelli riguardanti la complessità dei testi forniti dagli altri software, col vantaggio di ottenere dettagli più precisi su certe categorie-chiave (il turpiloquio, per esempio, che rappresenta un tabù nei testi formali come in quelli giornalistici veritieri).

La gran parte delle analisi condotte sugli articoli è stata, quindi, quantitativa: le osservazioni che abbiamo compiuto riguardano principalmente i dati ottenuti in maniera automatica a partire da una grande quantità di articoli. Solo nel caso dei titoli, analizzati separatamente, abbiamo optato per un'analisi qualitativa: siccome sono più brevi e hanno caratteristiche molto ricorrenti, è stato

possibile leggerli tutti e arrivare a conclusioni precise sulle differenze tra le bufale e gli articoli veritieri. Inoltre, su un insieme così ridotto di dati linguistici, l'analisi statistica non sarebbe risultata affidabile.

Creare dei corpora di testi e ottenere i dati tramite software, nonostante la notevole mole di dati e lavoro che richiede, è solo l'anteprima dell'effettivo lavoro di ricerca. La vera sfida, una volta ottenuti i dati, è dare un senso ai numeri. Senza scendere eccessivamente nei dettagli, i paragrafi che seguono si concentrano su alcune delle principali conclusioni che si possono trarre dall'analisi.

2.2.4. I campi semantici

I dati sul lessico ottenuti tramite *TalTac*[2] si possono collocare con facilità in tabelle, che permettono di mettere a confronto le forme o i segmenti presenti in notizie veritiere e false. In questo modo possiamo individuare sovrapposizioni e divergenze nei campi semantici (gli insiemi di parole o espressioni che riguardano un certo argomento): per esempio, dalla presenza di *Schettino*, *Costa*, *Giglio* in uno solo dei corpora potremmo notare che la tragedia del naufragio della nave di *Costa Crociere* avvenuto nel 2012 non viene mai menzionata negli altri corpora; o, alla luce della ricchezza di lessico specifico riguardante economia e finanza (*BOT*, *spread*, *Euribor*, *deficit*, *tassi d'interesse*, *Bankitalia*), rilevare che gli autori di determinati articoli si occupano spesso della questione del debito italiano.

Ciò che accomuna tutti i corpora sono gli aspetti più generali dell'attualità: per esempio, *tasse* compare 5 volte in *Bufale* e 6 in *Repubblica*, *economia* 16 volte in *Repubblica* e 14 volte in *Lercio*. A distinguere i corpora è invece la concentrazione di lessico di campi semantici diversi, che rivelano gli argomenti su cui insistono, nonostante la varietà di temi affrontati dall'insieme di articoli sia, come abbiamo visto in precedenza, presumibilmente molto simile.

In *Repubblica*, a emergere è la terminologia specifica della cronaca politica, economica e giudiziaria: se nelle notizie false questi argomenti vengono toccati almeno superficialmente, in *Repubblica* la concentrazione di lessico specifico lascia supporre che si tratti effettivamente del cuore del contenuto del quotidiano (un aspetto che viene confermato da precedenti ricerche sulla

lingua dei giornali). Alcune parole si ritrovano molto più spesso (per esempio, *per cento*, legato presumibilmente all'elencazione di dati economici, si ritrova 34 volte in *Repubblica* contro solo 3 in *Bufale*). Altre, anche se compaiono una sola volta in tutto il corpus (vengono definite tecnicamente "hapax legomena"), costituiscono comunque un segnale nel momento in cui formano un campo semantico numeroso (per esempio, espressioni che si riferiscono alla politica estera come *Osce*, *Eurozona*, *Bce*, *Casa Bianca*, *Angela Merkel*, *Alexis Tsipras*, *Sud Sudan*, *Nazioni Unite*, *Xi Jinping*, compaiono in *Repubblica* ma non in *Bufale*).

In *Bufale*, invece, sono i temi medici e scientifici a spopolare: oltre a un eloquente dato di 83 occorrenze contro 5 in *Repubblica* per *cancro*, molta terminologia medica risulta assente negli articoli veritieri (per esempio: *H1N1*, *anoressia*, *oncologica*, *diagnosticato*).

Per finire, in *Lercio* spicca una manciata di nomi di personalità della politica e dello spettacolo, su cui la penna irriverente degli autori sembra concentrarsi (*Salvini* compare 12 volte, *Adinolfi* 6 volte), mentre quasi nessun nome proprio compare una sola volta: personaggi politici o della cronaca quindi sembrano non essere citati occasionalmente a scopo informativo, ma piuttosto come vittime di una satira ripetuta. Non mancano turpiloquio ed espressioni legate alla sfera sessuale o politicamente scorrette (10 occorrenze per *sperma*, 5 per *cazzo*, 6 occorrenze per *Gesù*, tre occorrenze per *negri*). Fa capolino anche il linguaggio colloquiale, regionale e creativo (*figo*, *umarell*, *senzarotella*): parole che, per quanto inoffensive, riserviamo normalmente alla conversazione informale e non troviamo quasi mai in un articolo pubblicato su un quotidiano autorevole.

2.2.5. Attualità e atemporalità

Il lessico delle notizie veritiere e false contiene diversi indizi relativi all'effettivo legame con ciò che accade in Italia e nel mondo. Innanzitutto, i campi semantici possono suggerire che le notizie false si occupino solo superficialmente dell'attualità. Come abbiamo visto nel paragrafo precedente, pur confrontando corpora di articoli che apparentemente coprono gli stessi argomenti e nello stesso periodo di tempo, troviamo grandi differenze a livello di lessico e quindi presumibilmente di contenuti coperti dalle notizie: per esempio, *Repubblica* sembra darci molte più informazioni sulla politica economica di quanto non accada in *Bufale*.

Altri strumenti, come l'indice "con informazioni" di *Corrige!*, sostengono l'ipotesi. Questo indicatore individua le parole legate a quelle che vengono definite *conoscenze enciclopediche* del lettore: tutto un insieme di informazioni specifiche, come nomi geografici, nomi propri di personalità note della politica e dello spettacolo, organizzazioni internazionali, che da un lato devono essere presenti per collocare una notizia nel contesto dell'attualità, e dall'altro implicano che il lettore ne sia almeno un po' informato e che quindi possa capire autonomamente le relazioni che collegano diversi fatti. Questa misurazione vede *Repubblica* molto più ricca d'informazioni specifiche (il 60% in più che in *Bufale*), mentre *Lercio* si colloca in posizione intermedia (35% in più rispetto a *Bufale*).

L'analisi del lessico rivela anche che nelle bufale scarseggiano le espressioni di riferimento al trascorrere del tempo, definite in linguistica "deittici temporali": per esempio, *ieri* compare 49 volte in *Repubblica* ma solo 14 volte in *Lercio* e 6 in *Bufale*. Potrebbe significare che *Repubblica* parla di fatti ben inseriti nell'attualità, mentre nelle notizie false il legame è più vago, tanto che non serve (o è meglio evitare di) riferirsi al contesto della cronaca; le notizie false devono sembrare attendibili a lettori che le leggono in momenti diversi e per questo motivo devono essere scritte con pochi riferimenti all'attualità.

L'insieme di queste misure, unito all'osservazione di come le notizie false abbiano spesso vita (digitale) breve (i siti che le ospitano scompaiono periodicamente o comunque le cancellano) porta a sospettare che gli articoli falsi vengano costruiti a partire da strutture-canovaccio atemporali, a cui vengono cambiati di volta in volta pochi dettagli, mantenendo come costante il legame a tematiche sempre di sicuro successo quando si tratta di ottenere clic, condivisioni e *Mi piace*, come le tasse, la perdita di potere d'acquisto o il dilagare della criminalità (cfr. sotto § 3.2.1.). Se poi è vero che inventare storie ricche di dettagli è difficile e mette a rischio la plausibilità di una menzogna, dalla povertà di informazioni enciclopediche possiamo ipotizzare che gli autori di bufale preferiscano andare sul sicuro, o comunque prediligano la quantità di articoli prodotti (a partire da uno stesso schema) alla qualità (bufale ricche di particolari e plausibili anche per chi conosce bene l'attualità). Al contrario, la dovizia di dettagli (difficili da inventare in tale abbondanza) degli articoli di *Repubblica* conferma il loro stretto legame con l'attualità; gli studiosi della lingua dei

giornali parlano addirittura di contenuti al limite del comprensibile, tanto sono ricchi di particolari e difficili da decifrare da parte di un lettore occasionale.

2.2.6. Facile o difficile? Creativo!

Con l'analisi automatica dei testi, la complessità di articoli veritieri e falsi può essere individuata grazie ad alcuni indicatori che la valutano su diversi livelli dell'analisi linguistica: dal lessico (la maggiore o minore comprensibilità delle parole) alla morfologia (la forma più o meno complessa delle parole) alla sintassi (i legami tra le parole, che formano frasi più o meno lineari o contorte). Vista la ricchezza di dati, stabilire livelli precisi di difficoltà dei diversi tipi di notizie analizzati può aiutarci molto nell'individuare le bufale.

Il quadro che emerge, purtroppo, non è univoco. I valori della ricchezza lessicale (il rapporto tra numero delle parole diverse presenti in un corpus, contate ciascuna una sola volta, e il numero totale di tutte le parole presenti del corpus, comprese le ripetizioni) vedono *Lercio*, con 24,61%, a notevole distanza sia da *Repubblica* (22,02%) sia da *Bufale* (21,39%), il che indica un lessico più creativo nella testata parodistica, mentre il corpus di *Bufale* è quello in cui si ripete più spesso un numero minore di parole diverse. Anche la percentuale di hapax legomena (le parole presenti una sola volta) è un indice della creatività lessicale che conferma la posizione preminente di *Lercio* (63,21%), seguito, in posizione invertite, da *Bufale* (59,25%) e Repubblica (59,23%). Insomma, dal punto di viste del lessico, *Lercio* sembra avere uno stile più "mosso" e brillante degli altri due corpora.

La ricchezza lessicale, che fondamentalmente si basa sulla varietà delle forme usate, si scontra tuttavia con la comprensibilità del lessico. A questo proposito, il calcolo del *Vocabolario di Base* (VdB), cioè delle circa 7000 parole usate più di frequente nell'italiano moderno (e quindi presumibilmente comprensibili a tutti), indica che è invece *Repubblica* a usare il lessico più ricercato (15,78% delle parole non è presente nel VdB, contro il 14,50% di *Lercio* e il 14,04% di *Bufale*).

Un altro strumento di valutazione della difficoltà dei testi è l'indice GULPEASE, un calcolo della leggibilità basato sulla lunghezza di parole e frasi sviluppato dalla Sapienza - Università di

Roma: l'assunto è che frasi e parole più brevi rendano un testo più comprensibile. In questo caso è *Repubblica*, con un punteggio di 51 su 100, il corpus più leggibile, seguito da *Bufale* con 50 e *Lercio* con 49. Come si vede, in questo caso i tre corpora producono risultati molto simili: le notizie false usano lessico più semplice ma più variato e non risultano più facili da leggere rispetto agli articoli veritieri.

Passando alla morfosintassi, l'analisi tramite *TreeTagger* rivela che *Repubblica* è più ricca di nomi, il che conferma quello che già sappiamo dello stile giornalistico, più propenso a soluzioni che prevedono la nominalizzazione, ossia l'uso di un nome, più distaccato e formale, al posto di un verbo, più diretto e informale: un esempio che potremmo trovare sul web è "La *registrazione* al sito è avvenuta con successo" invece di "Sei riuscito a *registrarti* al sito". Sorprendentemente, però, *Repubblica* evidenzia una percentuale di aggettivi minore di *Lercio* e *Bufale* (mentre, normalmente, gli aggettivi si accompagnano ai nomi): è verosimile che questo dato rispecchi la preferenza, da parte delle notizie false, per l'aggettivazione sovrabbondante e iperbolica (per esempio *allucinante*, *agghiaccianti*, *scellerate*, *disastrosi*).

Per quanto riguarda i verbi, un'altra piccola sorpresa è la maggior frequenza di congiuntivo e condizionale nelle notizie false (rispettivamente, 2,89% e 2,50% nel corpus *Lercio+Bufale)* rispetto alla *Repubblica* (2,37% e 1,78%): si tratta di un utilizzo dei verbi tipico di testi formali e non scontato, alla luce degli altri dati sullo stile delle bufale. Forse ci saremmo attesi una minore formalità negli articoli ingannevoli, anche se abbiamo già avuto modo di notare la preferenza per alcuni tratti che caratterizzano un italiano più "corretto", quasi scolastico, come l'uso dei pronomi *egli, essi, costoro* ecc. (anche il passato remoto risulta più frequente: 2,26% nelle *Bufale* contro 1,40% in *Repubblica*). Tuttavia le notizie false evidenziano anche un uso più frequente dell'imperativo (1,47% contro 1,11% di *Repubblica*), che potrebbe indicare il tentativo di stabilire un contatto più diretto col lettore, confermato dall'uso più frequente della seconda persona nei verbi (per esempio, in *Bufale* troviamo 38 verbi che terminano in *-ete* ma solo 6 in *Repubblica*): per quanto riguarda la seconda persona, troviamo esempi come *Condividi questa vergogna italiana!* o *Controllate le bollette, vi vogliono fregare*.

2.2.7. Parla come mangi (seconda parte)

Infine, la differenza forse più evidente e trasversale a tutti i livelli di analisi dei corpora riguarda il registro linguistico. In generale, il registro di un testo viene collocato su una scala dal basso (informale, poco attento alla grammatica e alla struttura del testo) all'alto (formale, complesso, ricco di lessico ricercato). Tendenzialmente un registro si collega ai diversi usi della lingua (per scrivere un romanzo, per scrivere un articolo, per parlare a una conferenza, per spiegare una lezione, per chiacchierare...), e al rapporto tra gli interlocutori (useremo un registro informale con gli amici, formale con gli estranei). Per definire il registro occorre guardare all'assetto generale del testo, a tutti i livelli. Alcuni esempi saranno sicuramente più d'aiuto: l'italiano dei giornali evidenzia un registro medio (frasi brevi e informative, lessico specifico di tipo politico, economico ecc.), l'italiano della burocrazia un registro alto (lessico desueto, formale, frasi lunghe e complesse), l'italiano informale un registro basso (lessico quotidiano ed espressivo, poca attenzione per la correttezza grammaticale) e così via.

Repubblica dimostra, all'analisi, di rappresentare il modello di italiano giornalistico: i periodi (le frasi) sono brevi ma tendono alla nominalizzazione (cioè presentano più nomi che verbi) e, come avviene tipicamente nella lingua scritta, sono presenti in misura limitata solo mezzi linguistici usati per creare riferimenti interni al testo (*più avanti, prima, qui sotto* ecc.) e non alla situazione esterna (*qui, ora, noi, voi* ecc.). Ci sono pochi aggettivi dimostrativi (*questo* e i suoi derivati compaiono solo 161 volte in *Repubblica*, mentre li troviamo ben 365 volte in *Bufale*, con *Lercio* in posizione intermedia con 250 occorrenze). Si tratta di un fenomeno dovuto al fatto che chi scrive non ritiene necessario chiarire troppo spesso come si svolge il ragionamento (come se durante la lettura di un giornale qualcuno ci dicesse a ogni riga "Hai capito?", "Tutto chiaro?", "Ripetiamo di nuovo se ti sei perso qualcosa"): dopotutto, possiamo sempre rileggere. Nelle bufale, invece, gli autori sembrano ignorare deliberatamente questa norma dello scritto e forse "scrivono come parlano", un possibile risultato del tentativo di far passare il messaggio più efficacemente.

Lercio e *Bufale* mostrano tratti che avvicinano il loro stile a un registro più basso, influenzato dalla lingua parlata. La seconda persona viene usata più spesso, forse in funzione conativa (cioè

nell'intento di rivolgersi direttamente al lettore e persuaderlo di qualcosa, come nel linguaggio pubblicitario), e sfruttano maggiormente i riferimenti sia al testo che alla situazione esterna. Diversi campi lessicali-spia confermano l'ipotesi del registro informale: il turpiloquio e il lessico colloquiale e creativo in *Lercio*, i verbi procomplementari (tipici della lingua orale, caratterizzati dalla presenza fissa di particelle pronominali come *ci*, *la*, *ne* che hanno perso il loro significato originario: *starci, fregarsene*, *prendersela* ecc.) e le iperboli in *Bufale*.

In *Bufale*, inoltre, troviamo un curioso tentativo di elevare il registro tramite pronomi (*egli, ella*, *esso*, *essa*, *essi*) e dimostrativi (*ciò*, *tale*, *costoro*) percepiti come formali dal parlante inesperto ma in realtà ormai desueti e più vicini al linguaggio scolastico (solo gli insegnanti più tradizionalisti insistono ancora a imporre l'*egli* soggetto al posto del *lui*) o burocratico (pensiamo alle formulazioni da azzeccagarbugli che troviamo sui certificati rilasciati da uffici pubblici) che allo stile giornalistico.

Anche i campi semantici legati ai temi-cardine dei quotidiani contribuiscono a farci sospettare che l'imitazione del giornalismo autorevole azzardata dagli autori di bufale e parodie non sia poi così ben riuscita: come abbiamo visto, *Repubblica* coincide con le aspettative di un giornalismo concentrato su cronaca politica, economica e giudiziaria, mentre *Bufale* e *Lercio* attingono da temi presumibilmente più 'virali', come pettegolezzi, cronaca scandalistica, (pseudo)scienza.

Insomma, a ben guardare c'è una certa differenza stilistica tra i corpora. Se *Repubblica* passa a pieni voti la verifica dello stile giornalistico, non si può dire lo stesso di *Lercio* e *Bufale*, che non riescono a celare dietro le loro strategie linguistiche l'intento parodistico o ingannevole dei loro autori: anche quando cercano di imitare gli articoli di attualità, la povertà di dettagli, la scelta di temi insoliti e la scarsa capacità di imitare l'italiano giornalistico rivelano che si tratta di notizie false.

2.3. Un approccio quali-quantitativo: i titoli

I titoli sono un condensato di ciò che ritroviamo negli articoli: imparando a riconoscerne le particolarità, possiamo rafforzare le ipotesi fatte in base all'analisi degli articoli. Per via della loro brevità, tuttavia, le loro caratteristiche sono troppo diverse da

quelle del corpo degli articoli per poterli sottoporre all'analisi quantitativa. Inoltre, non sarebbe scientificamente accurato limitarsi a quest'ultima, dato che i titoli sono corti e, senza corpora abbastanza sostanziosi da far analizzare ai software, si rischiano risultati inaffidabili. Pur sfruttando alcune misurazioni, abbiamo dunque valutato i titoli con un'analisi qualitativa, cioè leggendoli a uno a uno, per poter ricavare qualche osservazione interessante.

Alcune delle particolarità che emergono, soprattutto in riferimento alla lunghezza, riguardano entrambi i corpora di notizie false, mentre i dettagli stilistici sono propri soprattutto di *Bufale*. I paragrafi che troverete qui di seguito possono quindi essere utili soprattutto per individuare notizie false ingannevoli.

2.3.1. Caratteristiche fondamentali

Ancor prima di leggerli, possiamo notare che i titoli di notizie veritiere e false si differenziano per la loro lunghezza media: i titoli di *Bufale* e *Lercio* sono lunghi più del doppio (13 parole) di quelli di *Repubblica* (5 parole). Riportiamo sotto un tipico esempio scelto dal corpus *Repubblica*.

> 2) Mogherini all'Onu: ecco il piano Ue sui migranti (9 parole)

Nelle bufale troviamo invece titoli come il seguente:

> 3) Lira italiana: è ufficiale, da Gennaio 2016 sarà reintrodotta la valuta italiana rimossa nel 2002 (15 parole)

Si possono dare due spiegazioni a questa differenza. La prima è che in un quotidiano come *Repubblica* normalmente viene usato un paratesto più complesso, formato da titolo, sottotitolo, occhiello, citazioni estratte dal corpo dell'articolo: le informazioni da dare al lettore possono essere diluite in tutte queste diverse componenti, perciò non è necessario che il titolo sia particolarmente lungo. Un'altra motivazione può essere dovuta alla piattaforma su cui vengono pubblicati gli articoli, che nel caso di *Repubblica* è sia digitale sia cartacea, con le limitazioni tipografiche che ne conseguono, mentre nel caso delle notizie false è solo online. Si potrebbe dire che i titoli di notizie false sono "nativi digitali" e non hanno limitazioni di spazio.

L'altra differenza fondamentale si riscontra nella struttura dei titoli. Su *Repubblica* i titoli seguono in massima parte poche strutture definite e ottimizzate per concentrare le informazioni in poco spazio, rimuovendo tutte le parole non necessarie. Un tipico esempio è una sola parola per introdurre l'argomento principale (*EU*, oppure *Elezioni*), separata con un segno d'interpunzione (i due punti o la virgola) dal resto del titolo, dove tutte le parole superflue vengono eliminate (nell'esempio che segue sono inserite tra parentesi quadre parole utili a un possibile "completamento" del titolo):

4) Is: 5 [terroristi sono stati] arrestati in Australia, volevano decapitare [dei] poliziotti

Nelle notizie di *Lercio* si imita questo stile, usando però più parole. Nelle bufale vere e proprie, invece, la differenza è evidente: i titoli sono più lunghi, lo stile è molto più libero, e compare spesso un riferimento diretto al lettore. Ecco un titolo di ben 30 parole:

5) Quello che non Vi dicono: ARGENTO COLLOIDALE meglio di qualunque antibiotico, potrebbe essere efficace anche contro Ebola. Ma non ce lo diranno mai perché non conviene alle lobby farmaceutiche !!

2.3.2. Ortografia e punteggiatura: niente penna rossa!

Se diamo un'occhiata più approfondita ai titoli di *Bufale*, salta all'occhio una certa libertà nell'ortografia. Gli errori veri e propri non sono così frequenti, ma in compenso emergono molti altri segnali di scarso controllo o poca attenzione alla forma. Tornando al lungo titolo citato nel paragrafo precedente, possiamo notare un uso delle maiuscole insolito (*quello che non **Vi** dicono: **ARGENTO COLLOIDALE** [...]*) e due punti esclamativi in chiusura, oltretutto separati dall'ultima parola con uno spazio. Tanto per avere un confronto, la Tabella 3 riporta la frequenza di punti interrogativi ed esclamativi nei titoli dei diversi corpora.

	Bufale	Lercio	Repubblica
Punto esclamativo	27	13	0
Punto interrogativo	11	5	2

Tabella 3 – Frequenza di punti interrogativi ed esclamativi nei titoli dei diversi corpora

Nei titoli di *Repubblica* l'esclamativo non viene usato nemmeno una volta, e gli interrogativi compaiono meno di un quinto delle volte rispetto a quanto accade in *Bufale*; *Lercio* sta a metà tra i due. Insomma, uno o addirittura due punti esclamativi o interrogativi sono un segnale che abbiamo scovato una bufala. Se poi troviamo qualche spazio di troppo (o mancante: nella stampa ufficiale l'uso della punteggiatura e della spaziatura è molto regolato), possiamo starne quasi certi.

2.3.3. Allusioni e clickbait: notizie o pubblicità?

Molti titoli di bufale si concentrano sulla reazione del lettore invece che sul fatto da descrivere: un tipico titolo di una bufala (in cui, peraltro, si può anche notare la mancanza di spazi dopo i punti interrogativi e fermi) potrebbe essere:

> 6) Sapete cos'è questo?A breve tutti noi ne avremo uno a vita.Alcuni già lo hanno.Ecco perchè.VIDEO

Il titolo inizia con un verbo alla seconda persona, che include il lettore nella notizia, e prosegue alludendo a un oggetto (presumibilmente ritratto in una foto a corredo) che però non viene descritto nel titolo, così come viene lasciato in sospeso il motivo del suo utilizzo. Questa tecnica, mutuata dallo stile giornalistico anglosassone e dal linguaggio pubblicitario, viene definita *clickbait* (letteralmente in inglese, "esca per i clic"): una frase accattivante che si rivolge direttamente al lettore e lo spinge a cliccare sul link per continuare a leggere e conoscere l'argomento.

2.3.4. Il lessico: è ufficiale, ma anche pazzesco

Abbiamo già notato la tendenza all'esagerazione durante la nostra analisi quantitativa. È una caratteristica, questa, che si fa ancora più evidente quando si tratta dei titoli: abbondano aggettivi come *incredibile, assurdo*, *shock*. Per gli autori di bufale sembra fondamentale stupirci e provocarci, iniziando dal titolo:

> 7) Secondo uno studio shock le mammografie sono una crudele bufala medica

Un altro aspetto curioso del lessico è la tendenza a introdurre il titolo con formule come *è ufficiale* (che compare in ben sei titoli su 150 di Bufale), *è stato stabilito* o simili. In retorica tali costruzioni, definite *apodittiche*, individuano gli strumenti linguistici che puntano a convincere il lettore che ciò che sta per leggere è proprio vero, per quanto improbabile:

> 8) Sbarchi: il Viminale ha deciso, dal 2015 ogni Italiano dovrà ospitare un immigrato per 30 giorni nella propria casa

Un'altra caratteristica, già individuata nell'analisi quantitativa, è l'inopportuna presenza di abbreviazioni:

> 9) ULTIM'ORA Sgarbi distrugge collezione opere d'arte Ricoverato alla neuro

Usi come quello nell'esempio (*alla neuro* è un'abbreviazione adatta al linguaggio colloquiale, ma inopportuna per un titolo di giornale) non comparirebbero nei titoli di un quotidiano autorevole, dal quale ci si aspetta un uso della lingua più formale.