Fake News Analysis: features exploration

December 6, 2018

1 Introduzione

In questo documento verrà affrontata la tematica dell'individuazione delle Fake News, con particolare attenzione rivolta all'analisi lessicale e di come sia possibile discriminare le notizie vere da quelle false basandosi su tecniche di Natural Language Processing.

Il documento presenta, nella prima parte, un'analisi esplorativa delle **"features"**: caratteristiche estraibili dalle notizie che possono essere sfruttate per distinguere quelle vere dalle false. Ne sono presentate diverse e, per ciascuna, viene evidenziato se si tratta di una caratteristica discriminante o meno. Nella seconda parte, alcune di queste features vengono utilizzate isolare alcune Fake News e si procede ad analizzarne manualmente il testo per cercare ulteriori indicatori o eventuali anomalie presenti.

2 Fake News

Come ormai noto e documentato, le Fake News sono state utilizzate per influenzare l'opinione pubblica, per manipolare il risultato di elezioni politiche e in generale per modificare il comportamento delle persone al fine di favorire, in qualche modo, gli interessi di chi le mette in circolazione.

Negli ultimi anni si è cercato di migliorare lo sviluppo di applicazioni che potessero riconoscerle e segnalarle il più tempestivamente possibile, direttamente sui social o grazie a plugin per browser. La tecnologia sembra essere ormai matura per permettere una classificazione molto precisa (alcune applicazioni vantano una precisione dell'80%, o addirittura 95%) [1] e, anche se sono già disponibili dei software, le fake news continuano a proliferare.

Sono state pubblicate ricerche e sondaggi, [2] [3] sono stati resi disponibili una grande quantità di dataset già classificati e si tengono periodicamente dei contest pubblici in cui gli sviluppatori propongono soluzioni sempre più precise; i dati e la tecnologia permettono ormai a chiunque di esplorare nuove soluzioni.

2.1 Individuazione

Il concetto principale da sviluppare per riconoscere una fake news è individuare delle caratteristiche peculiari che potrebbero, in maniera molto generale, appartenere alla categoria delle notizie false ma non appartenere a quelle vere (o viceversa). Se si volesse costruire un classificatore sarebbe necessario, in fase preliminare, definire appunto su quali caratteristiche basare la distinzione: per esempio, se si scoprisse che gran parte delle fake news utilizza un certo tipo di lessico oppure è formata da frasi molto (o molto poco) complesse, allora si potrebbe sviluppare un software che estragga queste caratteristiche dalle notizie e, per via di un training appropriato, che riesca a classificarle nella giusta categoria a seconda della caratteristica considerata. Non è purtroppo un compito immediato identificare una qualità discriminante, che potrebbe risultare valida soltanto per alcuni sottoinsiemi delle notizie considerate e che comunque si potrebbe rivelare non più valida col passare del tempo; in questo documento verranno esplorate alcune di queste caratteristiche per cercare di scoprire, per quanto possibile, se si tratta di peculiarità che possono aiutare nella classificazione.

3 Features

Verrà utilizzato il termine **"feature"** per indicare le diverse caratteristiche pensate che possono aiutare a distinguere le notizie.

Sono state individuate inizialmente 3 feature, che si pensava potessero discriminare in parte le notizie vere dalle fake news: varietà lessicale, complessità morfologica e quantità di tweet correlati.

Quantità di tweet correlati: per ogni notizia sono state inserite le prime parole nella ricerca di Twitter, per poi contare quanti risultati ha prodotto la ricerca. La logica dietro a questa feature è cercare di capire se il titolo o l'inizio della notizia riscuote particolare successo sui social, ovvero se in generale c'è molta più discussione in-

torno a notizie non confermate rispetto a quelle vere. Per come è stata concepita la feature, però, risulta difficile individuare quali termini all'interno della notizia utilizzare per la ricerca, dato che l'uso di troppe parole non produce mai nessun risultato.

Complessità morfologica: analisi della struttura delle frasi che compongono la notizia e del loro albero di parsing, per determinare quanto e' complessa, quante frasi la compongono e come sono intrecciate fra loro. Si è ipotizzato che le notizie false potessero essere scritte in modo meno preciso, complesso, in quanto non redatte da professionisti del mestiere come giornalisti o scrittori; ma ci si potrebbe anche aspettare l'opposto: chi le scrive potrebbe essere costantemente aggiornato sui risultati ottenuti nel campo della classificazione delle notizie e comportarsi di conseguenza, adattando lo stile di scrittura per eludere i classificatori che fanno leva su alcune particolari caratteristiche della notizia.

Varietà lessicale: rappresenta il numero di vocaboli diversi utilizzati all'interno della notizia. Notizie molto brevi e con parole tutte diverse hanno quindi un valore molto alto per questa feature. Combinata con la complessità morfologica, fornisce una quadro abbastanza ampio dello stile di scrittura del redattore della notizia.

4 Implementazione

Il sistema e' stato implementato in modo da poter inserire, in qualsiasi momento, una nuova feature ed analizzare i risultati ottenuti; è anche possibile cambiare facilmente il dataset fornendo alcune informazioni per poterlo subito integrare.

Tutto il codice del progetto si trova su bitbucket: fake-news repository.

4.1 Dataset

Durante lo sviluppo si è utilizzato un dataset proposto da kaggle per un contest di machine learning riguardante la classificazione di fake news [4]. Il dataset presenta 20800 articoli caratterizzati da titolo, autore, testo della notizia ed etichetta (fake / non fake) e contiene notizie di vario genere, da articoli di giornale a tweet; le notizie etichettate come non fake provengono da fonti considerate affidabili, come famose testate giornalistiche. I record sono equamente distribuiti in 10413 fake e 10387 non fake.

È comunque possibile cambiare facilmente dataset ed eseguire il software in modo che produca i risultati per le feature che si vogliono considerare, applicate ai nuovi record, in modo che vengano prodotti nuovi grafici relativi ai nuovi dati.

4.2 Features Extraction

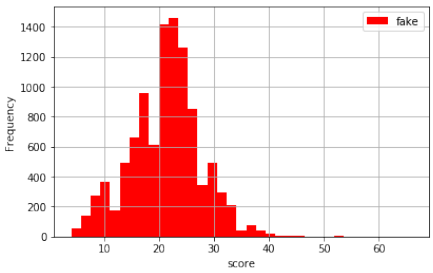
Di seguito verranno descritte dettagliatamente le features e la loro implementazione. Ciascuna di esse viene applicata a tutte le notizie del dataset e produce un valore numerico; sono poi proposti i risultati ottenuti sotto forma di grafici esplicativi. Per ogni feature, sono presenti due **istogrammi** che rappresentano la distribuzione dei valori della caratteristica sia per le notizie etichettate nel dataset come "affidabili" (non fake) che per quelle "non affidabili" (fake); è prodotto anche un

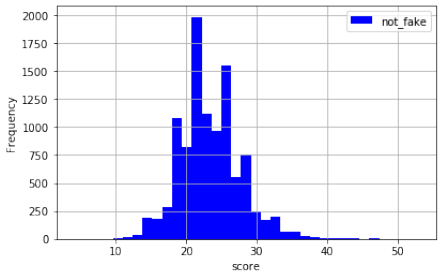
boxplot, per meglio analizzare la distribuzione dei risultati, sempre in duplice istanza per le notizie vere e false. Avendo i grafici divisi per le due classi di notizie, è facile capire se la feature pensata è un buon discriminatore: basta guardare se ci sono differenze percettibili nella distribuzione dei valori fra i due grafici.
Più avanti sono anche proposti degli **scatterplot**, per visualizzare se ci sono interdipendenze fra le features.

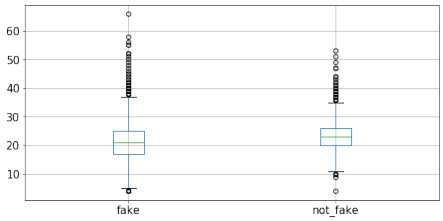
4.2.1 Morfological Complexity

Viene analizzata la morfologia della frase, ovvero la struttura grammaticale delle parole e la loro categorizzazione in nome, pronome, verbo, aggettivo. Sono evidenziati soprattutto i legami che questi elementi hanno fra loro all'interno della frase: più la struttura è complessa e più la frase risulta scritta in modo preciso e formale.
Per stimare, con un valore numerico, la complessità morfologica di una notizia, è stato utilizzato Stanford Parser [5]: analizza un testo e, per ogni frase, crea l'albero di parsing, individuando la struttura morfologica. Considerando poi la profondità dell'albero prodotto, si può valutare quanto la frase sia morfologicamente complessa.
Per ogni frase che compone la notizia viene calcolata questa profondità e poi si registra il valore massimo ottenuto in tutte le frasi: quello è il risultato della feature "morphological complexity".
Al parser viene dato in input tutto il testo della notizia, senza modifiche o preprocessing; l'unica modifica apportata riguarda la divisione dell'intero testo in frasi, necessario per utilizzare la funzione desiderata, ed è stata effettuata dividendo tramite un'espressione regolare che indica la punteggiatura di fine frase.

DISTRIBUZIONE FEATURE



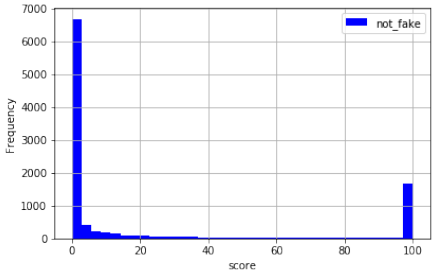


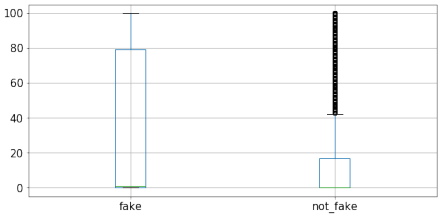


I primi due istogrammi rappresentano la distribuzione dei valori ottenuti applicando la metrica a tutte le notizie del dataset. Il primo riguarda soltanto le notizie etichettate come fake, il secondo quelle affidabili.
Entrambe le distribuzioni seguono un modello a campana, ma si possono notare delle differenze: nel primo caso, i valori sono più spostati verso sinistra (valori bassi più frequenti), mentre nel secondo sono più concentrati sul valore più frequente, con meno dispersione.
Questa prima osservazione potrebbe

aiutare ad isolare le notizie con valore di morphological complexity minore di 10 e classificarle come false.

Anche il boxplot evidenzia questa caratteristica: si osserva più dispersione nella distribuzione delle notizie fake, con una quantità maggiore di valori bassi rispetto a quelle non fake.

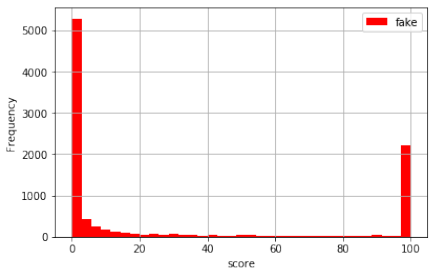




4.2.2 Twitter

Inizialmente il testo della notizia viene pre-processato, si rimuove punteggiatura e caratteri non validi e sono estratte le prime 5 parole che compaiono. Attraverso le API di Twitter si esegue una ricerca utilizzando questi primi termini e poi si contano i tweet correlati risultanti.
Il numero di risultati dipende molto da quali parole compaiono per prime nella notizia e soprattutto da QUANTI termini si sceglie di usare. Il numero massimo di post che Twitter restituisce è 100.

DISTRIBUZIONE FEATURE

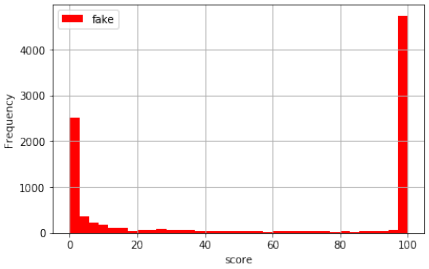


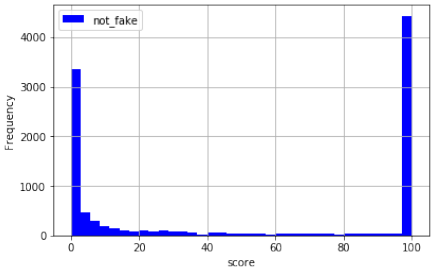
I grafici indicano che questa non è una buona feature da utilizzare per la distinzione: il numero di tweet correlati risultati è quasi sempre 0 oppure 100, senza una vera distribuzione a campana. Inoltre questo risultato è lo stesso sia per le fake news che per le notizie affidabili: non c'è modo di distinguere le due categorie basandosi soltanto sui grafici di questa feature.

Una spiegazione potrebbe essere che, se le prime parole scelte sono molto comuni o fanno riferimento ad un argomento molto discusso, il risultato è il massimo (100); se, invece, si tratta di parole poco usate il risultato scende subito a 0, senza valori intermedi. A causa della distribuzione atipica, il boxplot generato non riesce a distinguere fra i dati della distribuzione e gli outliers, che si trovano tutti nel mezzo, e risulta quindi poco esplicativo.

Per capire meglio se la forma del grafico prodotto dipende soltanto dai

parametri scelti, o se invece è a causa della feature in sè, combinata con la logica della ricerca di twitter, allora è stata provata una variante. Invece che usate le prime 5 parole per la ricerca, vengono prese le prime 3. Il risultato è il medesimo, con i valori più spostati verso il 100 (rispetto a prima, dove era il valore 0 a comparire più frequentemente). Non c'è una frequenza alta di valori intermedi, si ha sempre 0 oppure 100, in entrambe le classi di notizie.

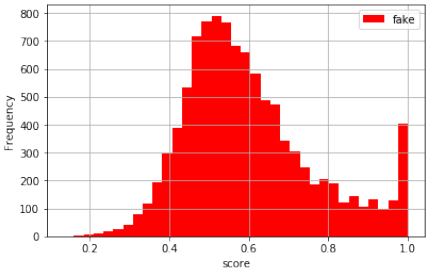


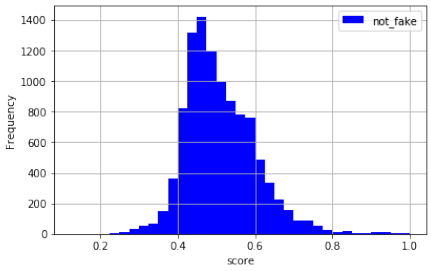


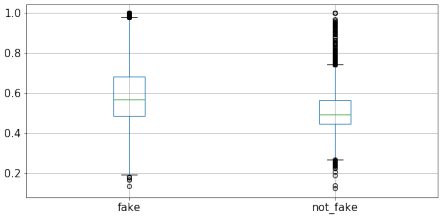
4.2.3 Lexical Variety

Per ogni termine diverso che compare nel testo della notizia, sono contate le sue occorrenze, utilizzando una funzione della libreria pattern.metrics. Il testo non viene manipolato: non è stata eseguita nessuna fase di preprocessing in cui le frasi vengono divise e le parole non sono sottoposte a stemming, per preservare la forma originale in cui è scritta la notizia.

DISTRIBUZIONE FEATURE

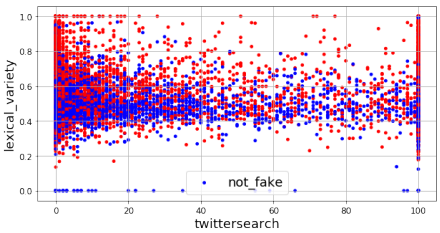






Come per morphological complexity, il grafico ha una forma a campana ed è ben distribuito. Si nota che, nel caso delle fake news, è presente una coda ampia oltre il valore 0.8, ci sono molti valori massimi e in generale la distribuzione è più ampia e non tutta concentrata al centro. Per quanto riguarda le notizie affidabili, queste sono quasi tutte concentrate verso 0.4 - 0.5.
Questo è forse il dato più significativo registrato: indica che le fake news hanno una probabilità maggiore di pre-

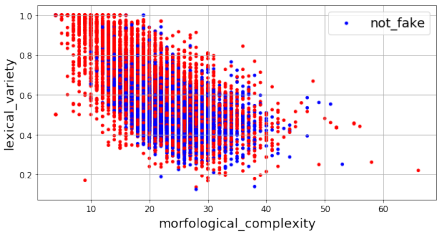
sentare una varietà lessicale più alta rispetto alle notizie affidabili, che si traduce nel fatto che il testo delle fake è molto più vario. Analizzando nello specifico la coda di valori a 1.0, si nota che sono tutte notizie di pochissime parole, tutte diverse.



La feature di Twitter combinata con le altre due non produce nessun risultato. Invece se si considerano le due caratteristiche linguistiche insieme, si nota qualcosa di più interessante.

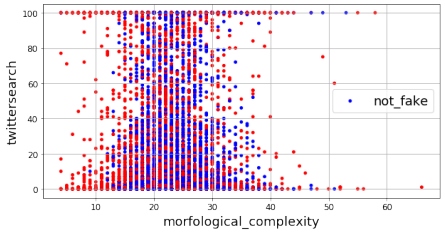
4.2.4 Combinazione Features

Invece che considerare soltanto una feature, è possibile analizzare i risultati di due (o più) caratteristiche prese insieme. Misurando l'interdipendenza fra due feature è possibile visualizzare se una delle due influenza l'altra; inoltre può servire per isolare delle regioni in cui sono presenti soltanto fake news e quindi registrare il range di valori di entrambe le feature in cui si ottiene questo isolamento.



È possibile isolare una zona in cui sono presenti soltanto fake news: queste hanno tutte un basso valore di complessità morfologica e un alto valore di varietà lessicale.
Questo risultato conferma quanto trovato in precedenza: nel dataset considerato, le notizie con poche parole tutte diverse, con una forma della frase molto semplice sono quasi sicuramente false.

Inizialmente sono riportati due scatterplot in cui vengono combinate due feature senza ottenere un risultato significativo: i valori sono sparsi in tutto il grafico e le notizie fake e quelle non fake sono mischiate fra loro; non è possibile trovare una zona in cui si raggruppa una categoria.



5 Conclusioni e Lavori futuri

È stato analizzato il dataset applicando diverse feature a tutte le notizie che conteneva e sono stati generati i grafici delle distribuzioni, per analizzare i valori prodotti dalle varie caratteristiche. Per le feature linguistiche si

è notato che all'interno di un certo range erano presenti soltanto fake news, quindi sono state considerate assieme per isolare in modo più preciso la zona contenente le notizie false. Utilizzando questo raggruppamento è stato possibile separare dal dataset le notizie false per analizzarle nel dettaglio.

Grazie all'analisi svolta è stato possibile classificare "manualmente" soltanto una parte delle fake news contenute nel dataset, ovvero quelle con bassa complessità morfologica e alta varietà dei termini. Proponendo nuove feature possibili e ripetendo l'analisi potrebbe essere possibile trovare altre caratteristiche che aiutino nella distinzione e, avendo a disposizione abbastanza indicatori, riuscire a distinguere interamente le notizie affidabili da quelle false.

References

[1] https://www.fakenewschallenge.org/

[2] Zubiaga, A., Aker, A., Bontcheva, K., Liakata, M., et Procter, R. (2017). Detection and resolution of rumours in social media: A survey.

[3] Conroy, N. J., Rubin, V. L., et Chen, Y. (2015). Automatic deception detection: Methods for finding fake news. Proceedings of the Association for Information Science and Technology, 52(1), 1-4.

[4] https://www.kaggle.com/c/fake-news/

[5] Sebastian Schuster and Christopher D. Manning. 2016. Enhanced English Universal Dependencies: An Improved Representation for Natural Language Understanding Tasks. In LREC 2016.