Progetto HPSA – Master SoBigData 2021

Elisa Mercanti, Gianluca Guidi, Marzia Longo, Vincenzo Aiello

1 INTRODUCTION AND OBJECTIVE

Il dataset utilizzato contiene una **raccolta di news** dal 2012 al 2018 (https://www.kaggle.com/rmisra/news-category-dataset). Nello specifico, per ogni news vengono riportati: il nome dell'autore, la categoria a cui appartiene, il titolo, una breve descrizione, il link e la data di pubblicazione. Il file json scaricato ha una dimensione di 80 M, contenente circa 200.000 records. L' **obiettivo** di questo lavoro è stato quello di costruire un predittore che permetta di classificare le news partendo dal titolo e dalla breve descrizione presente.

2 DATA UNDERSTANDING AND PREPARATION

In questa fase abbiamo studiato i valori e le features presenti. Come accennato, abbiamo in totale sei attributi che di seguito riportiamo:

- **Category**: variabile nominale che descrive la categoria di appartenenza di ogni news. Contiene 200.853 valori, di cui nessun valore mancante;
- **Authors**: riporta l'autore della news e contiene 200.853 valori di cui 36.620 mancanti;
- **Headline**: dato non strutturato di tipo testuale, contiene 200.853 valori di cui 6 mancanti;
- **Short_description**: dato non strutturato di tipo testuale, contiene 200.853 valori di cui 19.172 mancanti;
- **Date**: variabile ordinale identificativa della data di pubblicazione. Contiene 200.853 valori, nessun valore mancante;
- **Link**: link alla news, contiene 200.853 valori di cui nessun valore mancante.

Notiamo che i missing values sono circa lo 0.3 % dei valori totali [**FIG_1**, APPENDICE], presenti per lo più nelle features 'Authors' e 'Short_description'. Al fine di preparare il dataset per la classificazione valuteremo se sia più sensato sostituire tali valori o eliminare l'intero record avente il valore mancante. Facendo una prima considerazione, sicuramente per 'Short_description' sarà complicato poter sostituire tali valori, essendo un dato non strutturato di tipo testuale ed essendo specifica per ogni news riteniamo che la sostituzione non sia il modo adeguato di procedere. Per 'Authors', invece, vogliamo prima verificare quanto possa essere rilevante per la specificità di ogni news: se un autore scrive di un'unica categoria o di poche, potrebbe essere interessante utilizzare anche l'autore nel classificatore. Se non sarà ritenuto interessante per il task di classificazione non si porrà il problema di gestione dei missing values.

Al tal fine, abbiamo ritenuto interessante studiare come il numero di news di ogni categoria variasse nel tempo e come ciò potesse o meno influenzare la scrittura di uno specifico autore.

Nell'individuare la variazione delle categorie nell'arco temporale, ci siamo focalizzati sul capire quante categorie uniche possieda il dataset e quali tra esse siano maggiormente rappresentate. Abbiamo individuato 41 categorie differenti, aventi una distribuzione molto sbilanciata: alcune categorie sono molto rappresentate, come POLITICS che contiene più di 30.000 news, altre invece che contano circa 1000 news [**FIG_2**, APPENDICE].

Il numero di categorie risulta essere molto elevato per la classificazione e, inoltre, questo forte sbilanciamento della frequenza dei dati ci fa ipotizzare che la bontà della predizione potrebbe essere compromessa. Pertanto abbiamo ipotizzato di procedere in due differenti modi:

1) Ci è piaciuta l'idea di voler usare tutte le categorie e tutti i dati presenti nel dataset. Per questa ragione abbiamo deciso di creare una nuova feature, denominata '**Group Category**', ottenuta raggruppando le categorie per contenuto similare. In questa soluzione, ipotizziamo di avere un classificatore meno

specifico, che però avrà moda di allenarsi su molti più dati. Siamo riusciti ad individuare sei Macro Categorie [FIG_3, APPENDICE]:

- HEALTH&WELLNESS, 49475
- ENTERTAINMENT&GOSSIP, 49412
- POLICTS&EDUCATION, 33743
- INDIPENDENT VOICES, 28093
- TRAVEL&CULTURE, 24281
- SCIENCE&TECH&BUSINESS, 15849

Come si può notare anche qui vi è uno sbilanciamento delle Macro Categorie, ma risulta minore rispetto alla precedente.

2) Se per il task di classificazione la predizione non dovesse essere sufficientemente accurata, questo potrebbe essere dovuto al raggruppamento delle categorie. In questo caso, selezioneremo le top-6 categorie più rappresentative come target variable. Ciò significherà ridurre il numero di records utilizzati ma costruire un classificatore che debba riconoscere categorie pure, non aggregate.

Riteniamo interessante capire quale classificatore avrà la maggior accuratezza e valutare come il numero di record possa influire sulla predizione.

L'individuazione delle Macro Categorie è risultata utile anche nella visualizzazione della variazione di news nell'arco temporale. Nello specifico, per ogni anno abbiamo riportato il numero di news di ogni Macro Categoria e dal grafico emerge che [FIG_4, APPENDICE]:

- C'è stata una netta diminuzione delle news relative a HEALTH&WELLNESS;
- C'è un picco intorno al 2015 per la categoria ENTERTAINMENT&GOSSIP;
- Per POLITICS&EDUCATION vediamo la completa assenza prima del 2014, per poi assistere ad un notevole trend di crescita che raggiunge il picco nel 2017 per poi tornare a calare;
- C'è diminuzione costante anche per SCIENCE&TECH&BUSINESS;
- un trend abbastanza costante per le altre due categorie, TRAVEL&CULTURE e INDIPENTENT VOICES.

Abbiamo individuato delle Macro Categorie che risultavano avere delle variazioni più significative (HEALTH&WELLNESS, ENTERTAINMENT&GOSSIP, POLITICS&EDUCATION) e per queste abbiamo investigato quale categoria al loro interno fosse maggiormente responsabile di tale andamento.

Per la Macro Categoria HEALTH&WELLNESS, vediamo che la categoria WELLNESS risulta molto rappresentata solo nei primi tre anni mentre scompare del tutto negli ultimi tre. Stesso fenomeno risulta anche per HOME&LIVING, FOOD&DRINK, STYLE&BEAUTY. Le restanti Macro Categorie sembrano avere un leggero aumento a partire dal 2014. Ciò giustifica la diminuzione vista nel grafico precedente, poiché le categorie più rappresentative di tale gruppo si concentrano nei primi anni a discapito di quelle meno rappresentate [FIG_5, APPENDICE].

Per ENTERTAINMENT&GOSSIP, vediamo che la categoria maggiormente rappresentativa è ENTERTAINMENT che segue lo stesso andamento della Macro Categoria e che le altre sottocategorie mostrano avere quasi tutte un andamento ad essa similare o in costante crescita intorno al 2014-2016 [FIG_6, APPENDICE].

Per POLICTS&EDUCATION, vediamo che la categoria maggiormente rappresentativa è POLITICS come ci suggerivano anche le analisi precedenti. A dimostrazione del perché la Macro Categoria segue il suo stesso andamento [FIG_7, APPENDICE].

Analizzata la variazione di frequenza di ogni Macro Categoria nel tempo, abbiamo individuato i tre principali autori aventi il maggior numero di news: Lee Moran, Ron Dicker, Reuters Reuters. Per ognuno abbiamo analizzato la variazione della frequenza delle Macro Categorie nel tempo:

- Per il primo autore, Lee Moran, deduciamo che la Macro Categoria maggiormente trattata è ENTERTAINMENT&GOSSIP, in termini percentuali per tutti gli anni. Inoltre, nei primi anni l'autore si occupava principalmente di quell'unica Macro Categoria ma negli anni successivi si è evoluto,

seguendo anche i trend di mercato, fino a focalizzarsi anche su **POLITICS&EDUCATION** [FIG_8, APPENDICE];

- Per Ron Dicker vediamo che dal 2014 in poi anche lui ha scritto delle Macro Categorie che erano in continua crescita, come **ENTERTAINMENT&GOSSIP** e principalmente **HEALTH&WELLNESS**. Ha iniziato la sua carriera trattando poche tematiche differenti, per poi invece allargare la sua scrittura a quasi tutte le Macro Categorie, specializzandosi in quelle precedentemente individuate [FIG_9, APPENDICE].
- L'ultimo autore analizzato, Reuters, mostra essere molto attivo nei primi anni, con una preferenza per le Macro Categorie **HEALTH&WELLNESS** e **SCIENCE&TECH&BUSINESS.** Dal 2014 vediamo però un netto calo di pubblicazioni, confermando tuttavia le categorie trattate, fino al loro stesso appiattimento negli anni successivi [FIG_10, APPENDICE].

In generale possiamo dedurre che ogni autore ha sì una sua specifica categoria, ma sembra quasi sempre trattare anche altre. Notiamo un loro leggero allineamento a quelli che risultano essere i topic trend annuali e questo permette di affermare che 'Authors' non è una features utile per il task di classificazione.

Per la classificazione, oltre alla 'category' essenziali sono le features 'headline' e 'short_description'. Per entrambe è stato fatto un check di controllo sui duplicati: abbiamo cercato di individuare il numero di news aventi stessa 'short_description' e stessa 'headline'. Su 200853 news totali, **697** sembrano essere **duplicati**, circa lo **0,003 %**. Essendo tale dato molto basso e quindi ininfluente sulla bontà di predizione, abbiamo deciso di tenere i duplicati. Per i missing values presenti nelle due features, abbiamo deciso di rimuovere l'intero record non avendo altre alternative di sostituzione disponibile. In totale sono state rimosse 19718 records.

Per quanto riguarda le features utili al task di predizione, delle sei iniziali abbiamo ritenuto utili: 'Category', 'Headline', 'Short_description' e, ad esse, abbiamo aggiunto la 'Group Category'.

3 FEATURE EXTRACTION AND MODELLING

Come variabili nel task di classificazione abbiamo individuato:

- 'Group Category' come variabile target
- 'Headline' e 'short_description' come predittori.

Essendo i due predittori dati non strutturati di tipo testuale è stato necessario manipolare questi testi. Nello specifico è stato utile individuare i token presenti in essi, al fine di calcolare l'indice TF-IDF che rappresenta l'importanza di una parola nel documento o in una collezione di documenti.

Nello specifico abbiamo portato l'intero testo in lower-case, sostituito abbreviazioni presenti, eliminato eventuali link e doppi-spazi, sostituito caratteri accentati ed eliminato punteggiatura e stopwords.

Abbiamo poi unito le due colonne 'Headline' e 'Short_description' in modo tale da effettuare un'unica tokenizzazione. Ottenuti i tokens, abbiamo calcolato l'indice TF-IDF: questa operazione ci ha consentito di ottenere una colonna aggiuntiva contenente un vettore di indici per ogni news.

Per la variabile target, invece, essendo una variabile categoriale è stato necessario un'indicizzazione al fine di poterla utilizzare per la classificazione.

Nella fase successiva, abbiamo diviso il dataset in training set, utile all'apprendimento del classificatore, e test set, utile alla valutazione del predittore. Abbiamo effettuato un campionamento casuale, in percentuale 70% training set e 30% test set.

Avendo un problema di classificazione multi-classe e lavorando con dati testuali, ci è sembrato interessante utilizzare due classificatori molto diversi:

1) **Random Forest:** è un classificatore molto utilizzato per problemi di classificazione multi-classe; è un ensemble classifier, ovvero costruisce differenti decision-tree su diversi sub-set del campione: la

predizione finale sarà data dalla media delle classificazioni dei diversi decision-tree. Essendo uno strumento poco adatto per dati di tipo testuale è possibile trovare un classificatore piuttosto debole, specialmente se solo poche features sono significative. Tuttavia, considerando il fatto che abbiamo calcolato l'indice TF-IDF attraverso il quale abbiamo già effettuato una selezione delle features rilevanti, anche il Random Forest potrebbe risultare un buon classificatore.

2) **Naive Bayes**: è un classificatore bayesiano semplificato, ovvero assume l'indipendenza delle variabili (la presenza o l'assenza di un particolare attributo in un documento testuale non è correlata alla presenza o assenza di altri attributi). Si è visto che anche se spesso tale condizione di indipendenza risulta violata, il classificatore funziona bene comunque. Inoltre, è un algoritmo che funziona molto bene su dati di tipo testuale e, in generale, su grandi quantità di dati e di features.

4 RESULTS AND CONCLUSION

Nel calcolo del TF-IDF, la dimensione della Hash table è stata scelta in prima battuta calcolando il numero di parole distinte presenti nel predittore (111'755) e scegliendo la potenza di 2 che lo approssimasse al meglio (65'356). Tuttavia, poiché il numero di parole distinte è un numero molto grande, abbiamo deciso di ridurlo (a 4'096) per alleggerire il costo computazionale del modello una volta appurato che tale riduzione non influisse significativamente sull'accuratezza della predizione. I risultati di seguito esposti sono stati ottenuti da quest'ultimo parametro.

Dall'applicazione dei due classificatori è risultato che:

1) Il Random Forest Classifier risulta avere una pessima performance, con **accuratezza** di **0.31** e una **F1-measure** di **0.20** . Come si evince dalla confusion matrix in Fig.1, il predittore ha classificato la quasi totalità dei records come HEALTH&WELLNESS, ovvero la Macro Categoria più numerosa. Infatti, le due Macro Categorie con un minor numero di records (**TRAVEL&CULTURE**, **SCIENCE&TECH&BUSINESS)** non sono state classificate.

2) Il Naive Bayes Classifier, invece, risulta avere una buona performance, con **accuratezza** di **0.59** e una **F1-measure** di **0.59**. La confusion matrix in Fig.2 conferma l'accuratezza nettamente superiore al classificatore precedente, registrando il valore più alto nella categoria di POLITICS&EDUCATION (0.73). Questo potrebbe esser giustificato dal fatto che nonostante non sia la Macro Categoria più numerosa, è quella più "pura", in quanto è composta al 90% da news relative solamente a POLITICS.

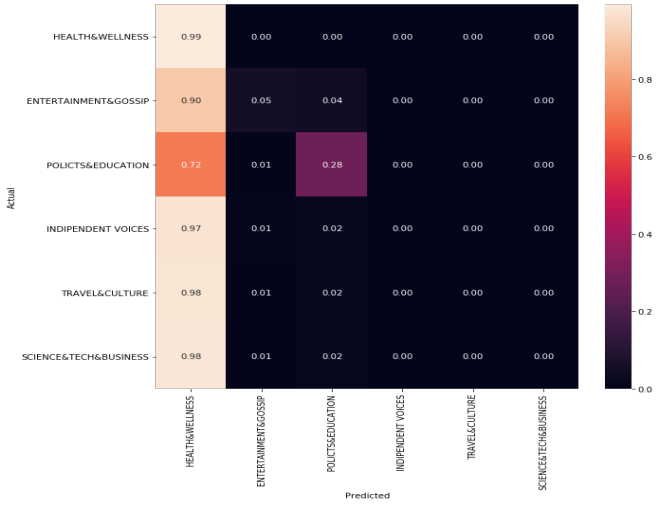

Fig. 1 Random Forest Confusion Matrix

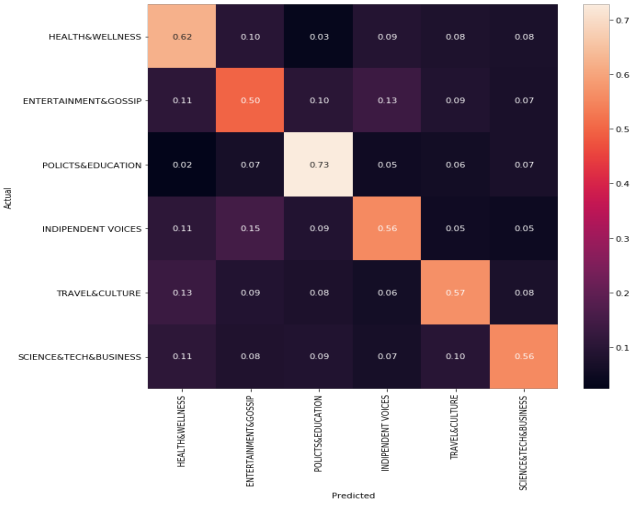

Fig. 2 Naïve Bayes Confusion Matrix

Dati i risultati ottenuti, abbiamo ritenuto interessante effettuare la stessa analisi non usando più come variabile target le Macro Categorie da noi ideate, ma le top-6 più frequenti categorie originali del dataset. Selezioniamo quindi i records delle categorie: 'POLITICS', 'WELLNESS', 'ENTERTAINMENT', 'STYLE & BEAUTY', 'TRAVEL' e 'PARENTING'.

Dall'applicazione dei due classificatori è risultato che:

1) Il Random Forest Classifier si conferma un classificatore poco adatto alle caratteristiche del nostro dataset. Infatti, riporta nuovamente le seguenti statistiche di predizione: **accuratezza** di 0.35 e una **F1-measure** di 0.19. Di nuovo, come evidenziato anche nella confusion matrix in Fig.3, il modello classifica tutte le news come POLITICS, che coincide con la categoria più numerosa.

2) Il Naive Bayes Classifier mostra invece una predizione ancora più accurata della precedente. Infatti, riporta valori di **accuratezza** e di **F1-measure** pari a 0.79.
La predizione risulta essere accurata ed omogenea per tutte le categorie(Fig.4), nonostante la disparità nel numero di records in esse contenuti. Ciò significa che a prescindere dalla frequenza di news, il predittore è capace di riconoscere in maniera precisa ognuna delle sei categorie.

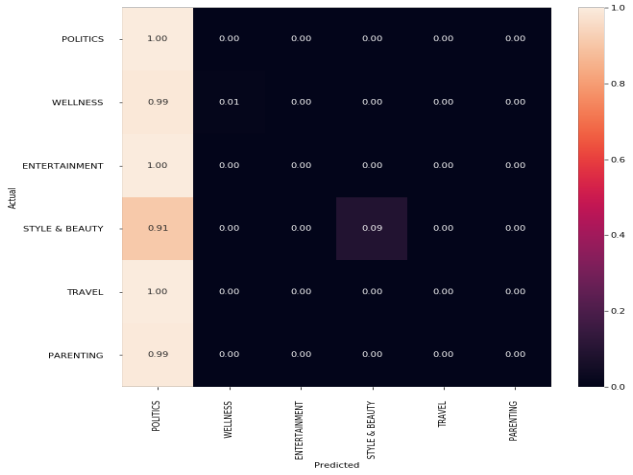

Fig. 3 Random Forest Confusion Matrix

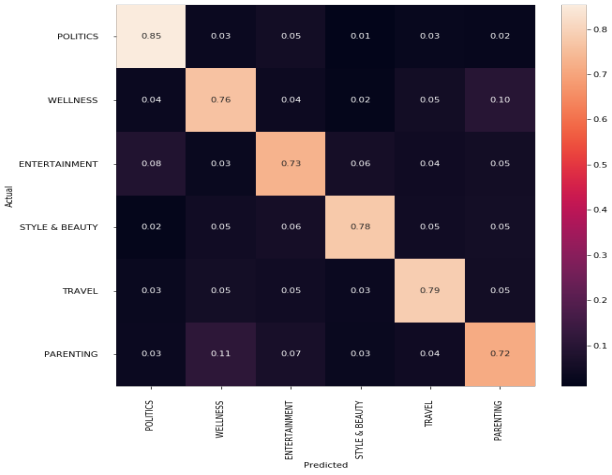

Fig. 4 Naïve Bayes Confusion Matrix

Alla luce dei risultati ottenuti con i nostri due classificatori, possiamo confermare che il Random Forest Classifier non è un classificatore adatto al tipo di analisi effettuata.

Nel secondo caso, con il Naive Bayes Classifier, abbiamo riscontrato l'adeguatezza di questo classificatore per questo tipo di analisi testuali.

Infine, data la migliore classificazione delle categorie originali rispetto a quella delle Macro Categorie da noi aggregate, osserviamo che nonostante un ridotto numero di records (e di categorie), avere delle classi più "pure" porta dei benefici alla performance del Naive Bayes Classifier che passa da 0.59 a 0.79.

5 SEZIONE AGGIUNTIVA – TAG CLOUD

Prima della classificazione vera e propria ci è sembrato interessate visualizzare le tag cloud per **Headline** e **Short_description** in modo da visualizzare le parole più frequenti presenti.
A tal fine abbiamo tokenizzato e lemmatizzato le parole presenti, creando due colonne aggiuntive aventi i token.

Headline

Short_description

Le abbiamo ottenuto con la libreria NLTK e wordcloud per la fase di visualizzazione.

Dalle tag-cloud vediamo che:

- in **Headline** le parole più frequenti sono **'photo'**, **'video'**, **'trump'**; parole che possono essere ricondotte alle categorie più rappresentative come **ENTERTAINMENT&GOSSIP** e **POLITICS&EDUCATION;**

- in **Short_description** abbiamo **'time'**, **'people'**, **'world'** e **'women'**; dalla descrizione risulta quindi esserci una maggiore difficoltà nell'individuare le categorie più rappresentative.

6 APPENDICE

FIG_1: descrizione per ogni categoria dei missing values.

FEATURES	MISSING VALUES	Tot VALUES	% di MISSING VALUES
Category	0	200853	0.000 %
Headline	6	200853	0.028 %
Authors	36620	200853	0.182 %
Link	0	200853	0.000 %
Short_description	19172	200853	0.095 %
Date	0	200853	0.000 %

FIG_2: frequenza di ogni categoria presente.

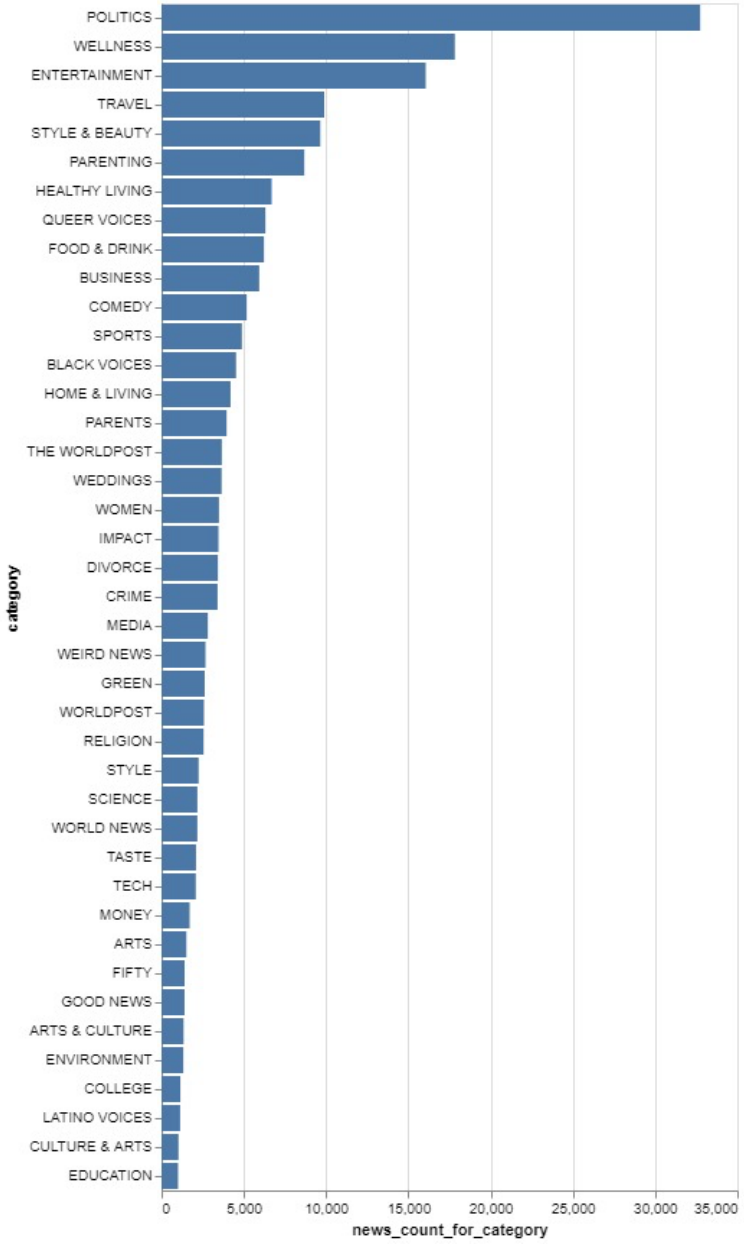

FIG_3: frequenza delle Macro Categorie.

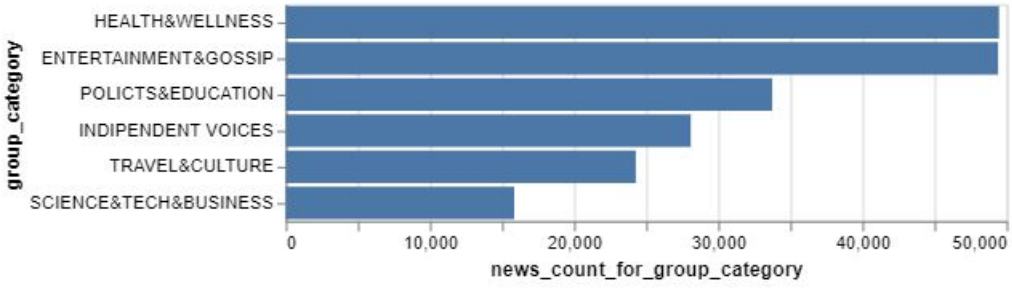

FIG_4: trend delle categorie presenti in ENTERTAINMENT&GOSSIP.

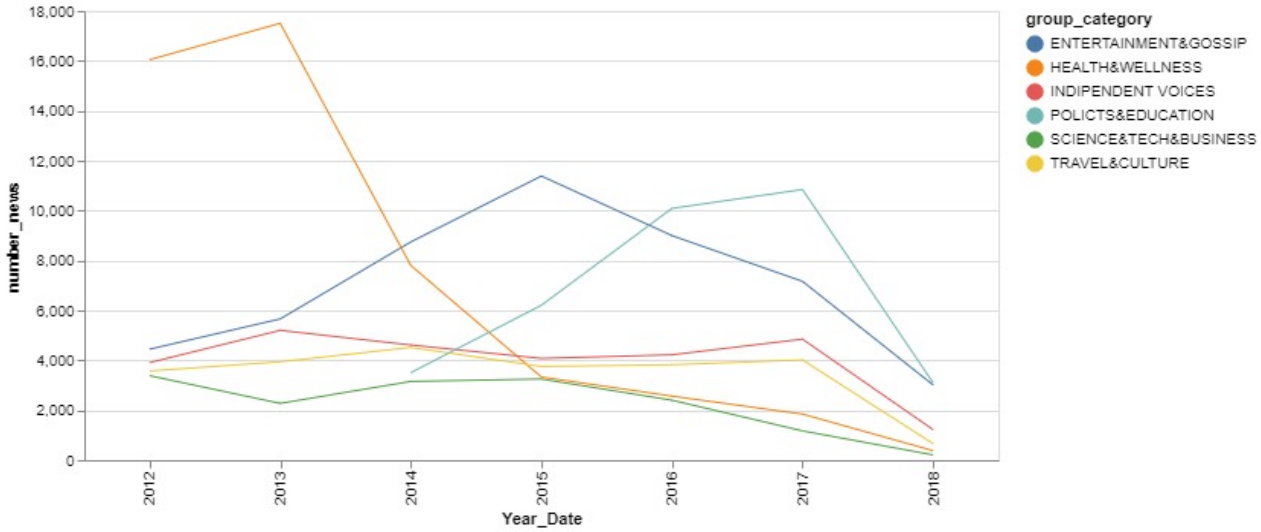

FIG_4: trend delle categorie presenti in HEALTH&WELLNESS.

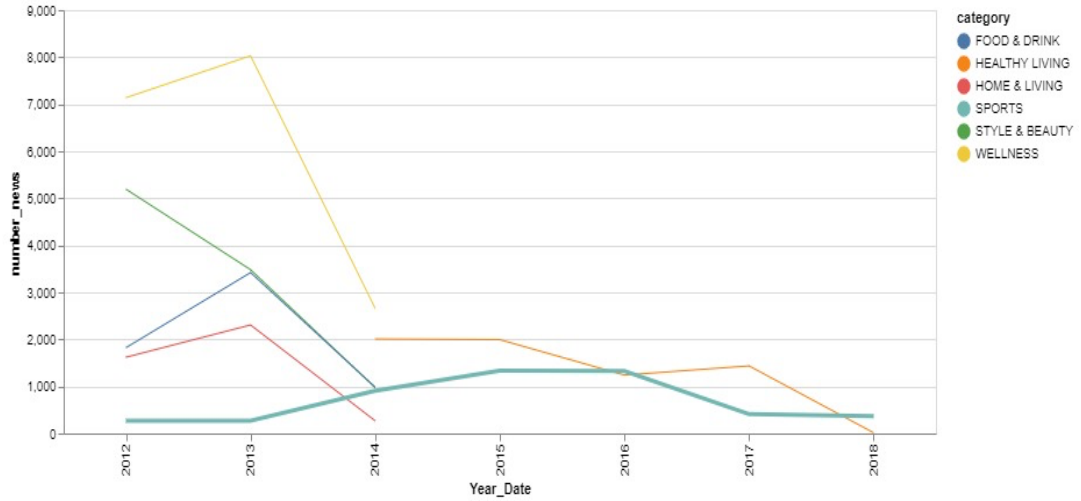

FIG_5: trend delle categorie presenti in ENTERTAINMENT&GOSSIP.

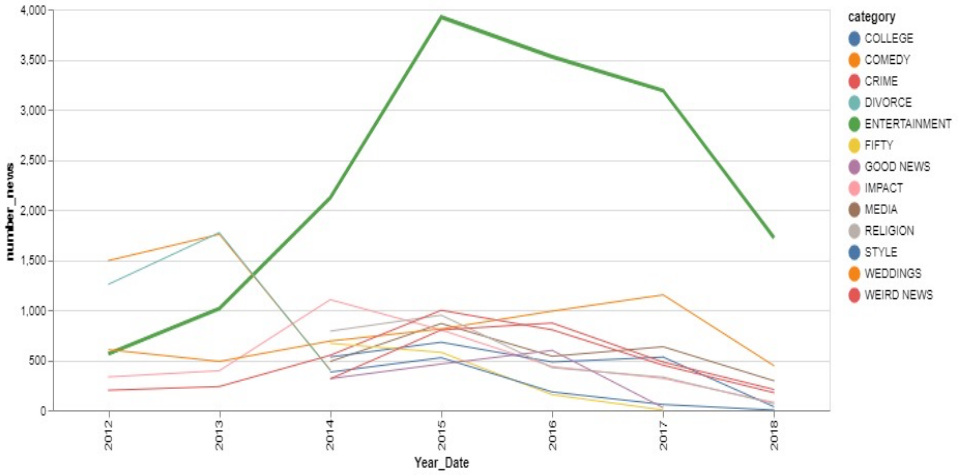

FIG_6: trend delle categorie presenti in POLICTS&EDUCATION.

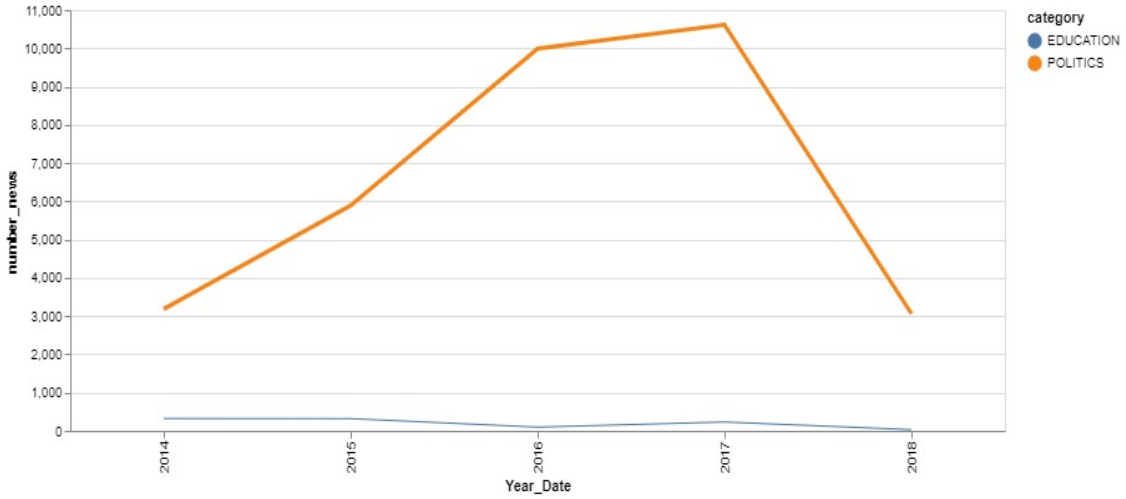

FIG_7: variazione delle Macro Categorie trattate nel tempo da Lee Moran.

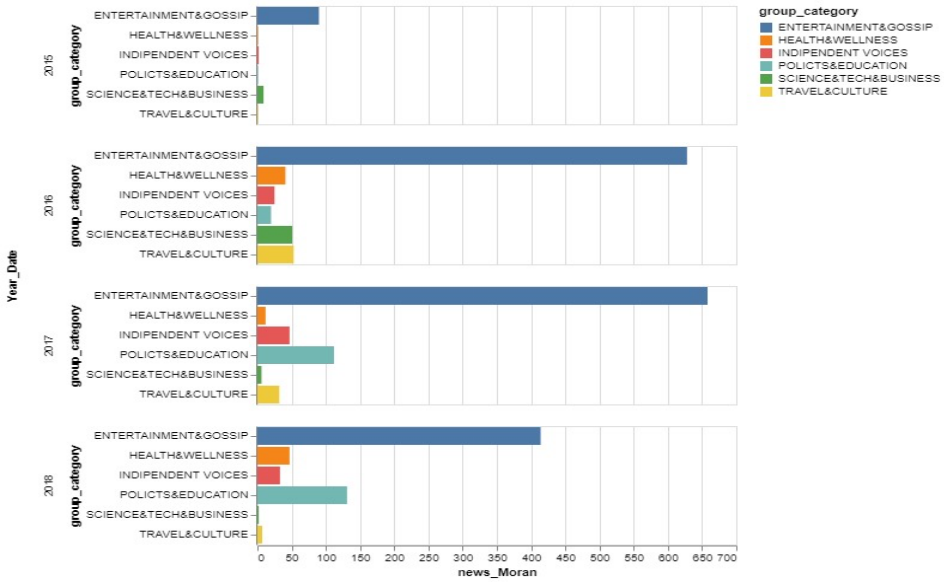

FIG_8: variazione delle Macro Categorie trattate nel tempo da Ron Dicker.

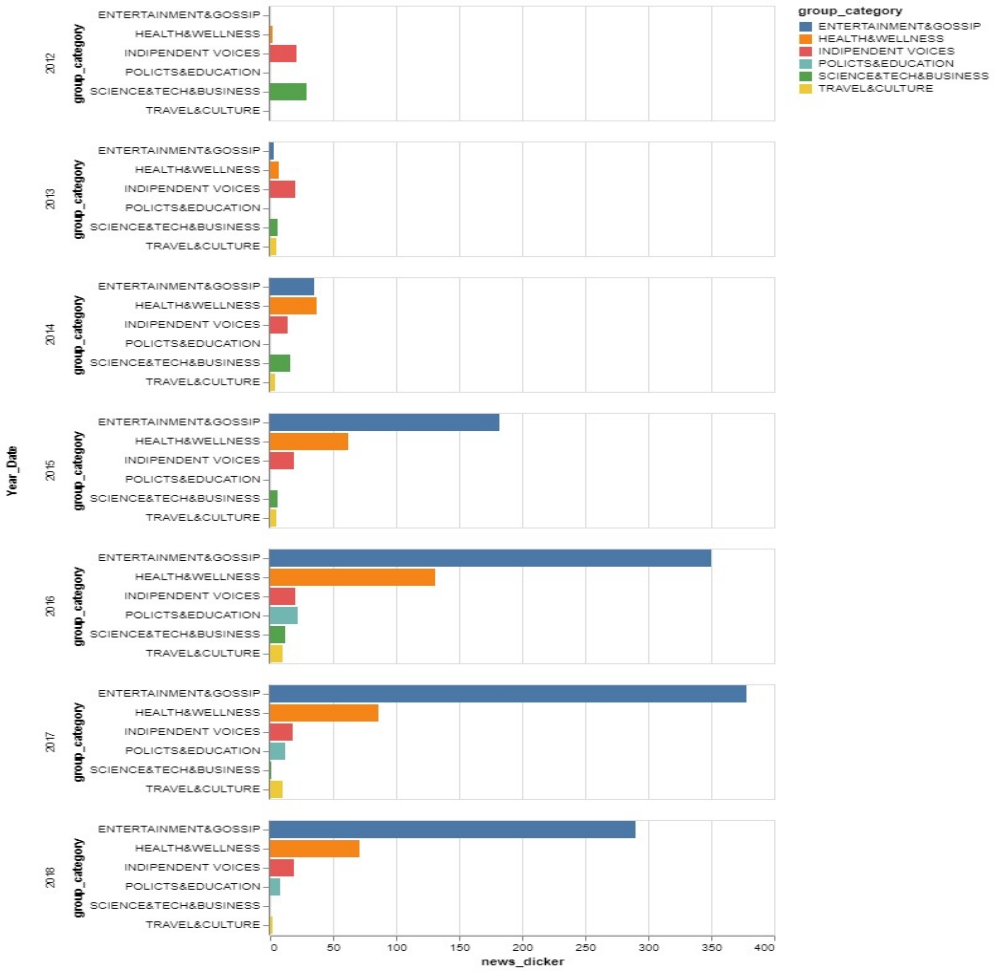

FIG_9: variazione delle Macro Categorie trattate nel tempo da Reuters Reuters.