Università degli studi di
Milano-Bicocca

Data Analytics

Final Project

BREXIT NEWS

Authors:
Gianmaria Balducci - 807141- g.balducci1@campus.unimib.it
Alessandro Guidi - 808865 - a.guidi@campus.unimib.it

July 3, 2021

1 Introduzione

Il significato della parola Brexit fa riferimento all'uscita del Regno Unito dall'Unione europea e deriva dalla crasi di due parole inglesi: Britain, *Gran Bretagna*, ed exit, *uscita*. Il termine è ormai entrato nel linguaggio di uso comune e negli ultimi anni, diventando nel tempo un sinonimo di profonda incertezza.

Il significato della Brexit però non è solo tecnico: come anticipato infatti l'uscita del Regno Unito dall'UE avrà conseguenze non solo sugli inglesi ma su tutta l'economia mondiale.

Il segnale lanciato dal popolo, prima con il referendum di giugno 2016 e poi ancora con le elezioni di dicembre 2019, ha messo in discussione il progetto Europa.

Per questo l'evento della Brexit è stato molto seguito e analizzato, in particolare con questo progetto si vuole analizzare gli articoli online scritti dai giornali su questo argomento sfruttando le tecniche di **Named-entity Recognition** per conoscere i maggiori protagonisti(persone,organizzazioni,luogi, etc..) più citati e le tecniche di **Sentiment Analysis** per valutare se gli articoli letti siano favorevoli o contro la Brexit.

2 Dataset

Il dataset utilizzato è diviso in 2 sezioni, una dedicata per il train e una per il test. Entrambi i file sono formati da soli 2 attributi(colonne), uno contiene un *URL* corrispondente all'articolo da esaminare, nel altro attributo è contenuta la data di pubblicazione dell'articolo. Gli articoli sono stati presi in un lasso di tempo corrispondente a quasi 5 mesi attorno alla data del **29/03/2017**, quando il Regno Unito ha presentato formalmente una notifica che riguardava la sua intenzione di uscire dall'Europa. In particolare il periodo preso in esame va dalla data del **30/11/2016** alla data **18/04/2017**. Il dataset relativo al train è un file *csv* contenente quasi 800000 righe(779207 articoli), il test invece ne contiene 10000.

Per il nostro progetto abbiamo scelto di tenere i file separati, iniziando a lavorare sul dataset train. Iniziando ad utilizzare le tecniche di **Scraping**,per ricavare le informazioni necessarie per fare un analisi del testo dal URL dato, molti articoli generavano degli errori tra cui: il testo dell'articolo vuoto o sostituito da un articolo diverso oppure il link dell'articolo non è più

raggiungibile. Inoltre la mole di dati data dal numero di articoli da analizzare , per le risorse a nostra disposizione è fin troppo elevata da elaborare.

3 Ambiente di sviluppo

3.1 Colab

Come visto a lezione, abbiamo deciso di usare il tool Colaboratory, conosciuto semplicemente come **Colab** per una maggiore famigliarità con il linguaggio di programmazione *Python* e per avere delle risorse Hardware maggiori rispetto a quelle disponibili da noi.

3.2 Librerie utilizzate

Per la gestione dei dataframe sono state utilizzate **numpy** e **pandas**, la fase di scraping è stata gestita utilizzando la libreria vista a lezione **newspaper3k**. Con **NTLK** è stata gestita la fase di preprocessing utilizzando le stopwords inglesi il work_tokenizer di questa libreria. Inoltre è stato utilizzato il pos tagger di NLTK per computare correttamente la lemmatization, usando WordNetLemmatizer di NLTK. Abbiamo utilizzato **Spacy** invece per la parte di nlp per l'aspect extraction, il lessico **Vader** con il SentimentIntensityAnalizer importato da **NLTK**.

Oltre a *Vader* abbiamo utilizzato un altra libreria con lo stesso scopo denominata come **Afinn**, per poterle confrontare.

Per i plot dei grafi è stata utilizzata la libreria **matplotlib** importando il modulo *pyplot*

Per poter analizzare gli URLs degli articoli, sono state utilizzate diverse librerie: **Urllib** per poter estrarre il dominio del URL e pulirlo dalle stinghe di query. Poi per poter associare ogni dominio trovato un Paese di appartenenza abbiamo usato la libreria **maxminddb-geolite2**, con la quale attraverso l'indirizzo IP, veniva restituito il Paese di appartenenza del server associato, per questa funzione è stato fornito supporto dalla libreria **socket**

4 Scraping

Il dataset utilizzato fornisce un campo contenente l'URL dell'articolo, da cui bisogna estrarre i dati dalla pagina web corrispondente, questa tecnica è conosciuta con il nome di **scapring**.

Abbiamo scelto di utillizare la libreria *newspaper3k* messa a disposizione su Python, in particolare il modulo *Article*. Da ogni link si ottiene un oggetto di tipo Article dal quale è possibile estrarre i dati dell'articolo, per poterlo leggere dobbiamo affidarci ai metodi della classe DOWNLOAD(), con il quale viene scaricato il codice HTML associato all'articolo. Un altro metodo che utilizziamo è PARSE() con il quale ricaviamo le informazioni che utilizzeremo per fare le nostre analisi, in particolare il testo dell'articolo e gli autori. Infine, con il metodo NLP(), sempre fornito dalla libreria citata, è stato possibile estrarre le parole chiavi all'interno dell'articolo, come entità e verbi.

C'è il rischio che ottenuto l'oggetto Article, nel momento in cui vengono utilizzati i metodi descritti precedentemente, vengono generate delle eccezioni laddove si è in presenza di un URL che non esiste più, a causa della rimozione dell'articolo o anche da un aggiornamento del sito il quale potrebbe aver cambiato i propri link.

Per questo sono stati utilizzate delle configurazioni adeguate a catturare queste eccezioni per non interrompere il programma. Anche non prendendo tutti i dati del dataset, ma 180 articoli per ogni giorno, questa tecnica richiede una grande capacità di Hardware, difatti l'algoritmo creato è stato in continuo aggiornamento per dividere la computazione totale e per permettere di arrivare ad un numero esiguo di articoli analizzati.

Vi sono dei giorni in cui il numero di articoli è minore di 180, questo può essere dovuto da 2 ragioni principali, la prima è gli articoli non sono riusciti a superare il nostro filtraggio di scraping, in quanto l'argomento trattato non riguardava la brexit. Il secondo motivo può dipendere dal dato dal fatto che molti articoli hanno generato le eccezioni di cui abbiamo parlato prima, così da diminuire drasticamente il numero di articoli.

```
topic = [["voted","leave","eu"],
         ["voted","leave","europe"],
         ["voted","leave","european union"],
         ["eu","uk","leave"],
         ["european","uk","leave"],
         ["labour party", "europe"],
         ["vote","referendum","eu"],
         ["european union", "vote"],
         ["uk","government","political"],
         ["exit","eu","voted"],
         ["exit","uk","voted"]]
```

Figure 1: Parole chiavi

Durante l'esecuzione dello scraping sono stati integrati dei controlli per filtrare gli articoli analizzati, in particolare abbiamo verificato che gli articoli parlassero davvero della *brexit*.

Per farlo abbiamo definito una lista di parole chiavi le quali, combinate tra loro, se venivano trovate all'interno del testo dell'articolo allora poteva essere analizzato e aggiunto al DataFrame che conterrà gli articoli che verranno analizzati. In aggiunta, abbiamo tenuto conto del titolo dell'articolo, nel caso in cui possa contenere la parola chiave del nostro progetto, "brexit", se vi era al suo interno, l'articolo veniva preso in considerazione. Gli articoli che non presentavano nessun titolo venivano scartati.

Inoltre è stato inserito un controllo in modo tale che il testo degli articoli siano tutti diversi tra di loro per evitare di avere delle ripetizioni.

5 Preprocessing

L'idea iniziale poter lavorare sul dataset è stata quella di ridurre il numero degli articoli, per farlo è stata fatta un analisi sugli eventi successi nel periodo che copre il database, così da individuare quali fossero i giorni più determinanti dell'evento della brexit.

Anche in questo caso però il numero restava elevato, in quanto in alcuni giorni si superavano i 5000 articoli pubblicati in solo 24 ore, inoltre una notizia potrebbe essere anche pubblicata a distanza di giorni rispetto ad un determinato evento, con maggiori approfondimenti o opinioni differenti.

Quindi abbiamo optato di filtrare gli articoli per giorno e per ogni giorno prendere lo stesso massimo numero di articoli, il numero scelto è stato di 180 articoli. Come detto precedentemente, vi sono alcuni giorni che presentano meno di 180 articoli dato dallo scraping effettuato.

Per raggruppare gli articoli abbiamo convertito la data nel dataset, che viene letta come una variabile stringa, in una variabile *Date* fornita dalla libreria di *pandas*. Dopo di che è stato aggiunto un attributo al dataset con il valore **date_f** , in cui veniva riportata solo la data e non l'ora della pubblicazione, in questo modo è stato possibile raggruppare i link per data di pubblicazione.

Per contare quanti articoli appartenessero ad una determinata data, è stato aggiunto un ulteriore campo al dataset, con il nome di **count**, il valore di ogni articolo è stato pari a 1, così che i valori degli articoli pubblicati lo stesso giorno si sommino per analizzarne il totale.

Raggruppando per giorno, sono usciti 140 gruppi, abbiamo notato che per un intera settimana, non sono stati inseriti articoli e quindi risultano vuoti, senza nessun articolo. Questi giorni sono stati eliminati.

Come detto in precedenza sono stati selezionati 180 articoli per ogni giorno, la scelta degli articoli è stata una selezione casuale ed enumerata, in quanto molti articoli durante lo **scraping** generassero delle eccezioni mentre altri restituivano un "corpus" vuoto, senza testo. Quindi creando i giusti controlli per gli articoli, per ogni data, si aggiungevano quelli idonei nel nuovo Dataframe fino a raggiungere la soglia prestabilita(o fino a che gli articoli pubblicati quel giorno finissero); arrivati al numero di articoli deciso, l'algoritmo passava alla data successiva.

Per ogni articolo idoneo da inserire nel nuovo dataset da studiare sono stati aggiunti ulteriori campi a quelli già presenti e aggiunti precedentemente. Il primo campo aggiunto **Article** contiene la referenza dell'oggetto data dallo

scraping effettuato. Gli altri campi sono informazioni con le quali è stato possibile reperire attraverso l'oggetto ritornato dallo scapring, in particolare: un campo aggiunto è stato quello degli autori denominato con **authors**, mentre un altro campo, quello più importante, è **text**, ovvero il testo dell'articolo. Infine sono stati inseriti altri 2 campi, il campo **title**, in cui viene fornito il titolo dell'articolo e il campo **keywords**, in cui venivano salvate le parole chiave citate nell'articolo esaminato.

5.1 Data Cleaning

Costruito il nuovo dataset contenente tutte le informazioni per effettuare l'analisi, lo step successivo è la pulizia dei testi degli articoli in modo tale da permettere agli algoritmi di NER Aspect Extraction e di Sentiment Analysis di estrarre tutte le informazioni necessarie, eliminando a priori i pattern del testo che non servono all'analisi e senza elementi che possano creare confusione nel testo. In particolare sono state applicate le seguenti operazioni:

- **Rimozione carattere di escape**
  In molti articoli sono presenti caratteri di escape più volte ripetuti, in questo caso con regex questi caratteri sono stati sostituiti con una spaziatura

- **Rimozione caratteri html**
  Nella fase di scraping ci sono stati alcuni errori nell'estrarre il testo di un articolo che possono portare alla presenza di caratteri html, in questo caso sono stati eli minati dal testo.

- **Rimozione url**
  Molti articoli presentano dei collegamenti ipertestuali ad altre pagine web ai fini dell'analisi non sono utili e sono stati eliminati.

- **Rimozione numeri**
  Abbiamo deciso di rimuovere i numeri dal testo in quanto nel task di aspect extracion non portavano contributi.

- **Gestione delle negazioni**
  Tutte le parole presenti in questa lista "don't", "not", "doesn't", "didn't", "wasn't", "hadn't" diventano not + token successivo. Per esempio don't like diventa not like

- **Tokenizzazione**
  Divisione del testo in token con **NLTK** word_tokenizer

- **Rimozione stop words**
  Dal testo sono state rimosse le stopword inglesi più l'aggiunta di caratteri speciali presenti spesso in questo corpus, che sono le più frequenti ma non portano significato all'analisi

- **Rimozione punteggiatura**
  Tutti i token del testo che appartengo a string.puncuation sono stati rimossi

- **Lemmatization**
  Task di normalizzazione del testo che riduce il token alla sua forma base a seconda del risultato del suo POS_TAG

6 Data Exploration

Il dataset non contiene label e dopo la fase di preprocessing avviene una fase di esplorazione del dataset creato, in particolare sul testo che sarà l'elemento centrale di questa analisi, inoltre sono stati presi in esame anche i campi come keywords e titolo dell'articolo. Per compiere questa esplorazione il testo è stato diviso in token per poter visualizzare i token più frequenti e capire se la fase di preprocessing ha dato dei risultati validi e comprensibili. Nell'immagine seguente è mostrata la wordcloud dei token più frequenti del testo. Il token più utilizzato in questo corpus è il verbo say, molto utilizzato

Figure 2: WordCloud articoli

nelle notizie per riportare le dichiarazioni che qualche persona, in questo contesto qualche politico, ha esposto riferendosi prevalentemente a token come UK, Brexit, people, government, may (riferito a Theresa May) ecc.. Da notare anche la presenza del token trump (Donald Trump) tra i più utilizzati ad indicare come la politica estera (in particolare gli U.S.A.) così come l'economia siano stati in qualche modo influenzati e abbiano influenzato il tortuoso processo della brexit. Per quanto riguarda i titoli delle notizie i token più frequenti come viene mostrato nella wordcloud seguente si concentrano sempre sull'evento Brexit, ma pongono più attenzione ai protagonisti politici e sull'accordo con l'europa.

Come possiamo vedere infatti saltano all'occhio token come Theresa, May, deal, plan, governement e successivamente sul voto e sui partiti *labour* è il Labour party schierato apertamente contro la brexit. Anche tra i titolo Donald Trump è spesso citato.

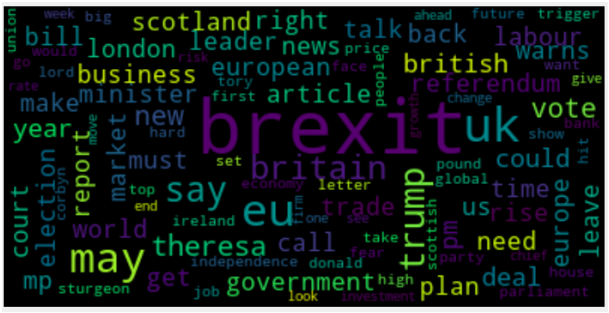

Figure 3: WordCloud titoli

Successivamente abbiamo applicato il task di Name Entity Recognition per capire quali sono le entità più presenti all'interno degli articoli

6.1 NER

Questo task è stato implementato utilizzando 3 diverse librerie.

Partiamo con la libreria **Spacy** che offre diverse funzioni e attributi per applicare la NER sul corpus di articoli. Ogni articolo che viene dato in input a spacy ha subito lo stesso preprocessing descritto prima tranne che l'operazione lower() che rende minuscoli tutti i token. Questo perchè il task di NER è facilitato quando vede i token che hanno un iniziale maiuscola e che quindi molto più probabilmente saranno taggati come NOUN nella fase di POS e i nomi propri spesso sono entità. In seguito vengono mostrate le entità più citate nel corpus e con esse le classi a cui appartengono sempre divise in ordine di cardinalità.

Entities

	name	count
0	EU	69916
1	Brexit	67056
2	UK	58407
3	Britain	34460
4	British	20529
5	Europe	17976
6	May	16243
7	US	14359
8	London	13797
9	first	13771
10	Trump	13489

Labels

	label	count
0	ORG	379114
1	GPE	323439
2	PERSON	306232
3	DATE	192631
4	NORP	117551
5	CARDINAL	56898
6	LOC	34552
7	ORDINAL	28530
8	WORK_OF_ART	12893
9	MONEY	11410
10	FAC	9413

Figure 4: Entities

Come ci aspettavamo le entità più citate sono **Brexit**, **EU**, **UK** che ci danno un anticipazione dei topic che andremo ad analizzare nellle fasi succesive. Da nitare la presenza di **May** che spacy classifica come date ma in realtà appartiene alla classe PERSON in quanto si riferisce al primo ministro di quel periodo **Theresa May** uno dei principali protagonisti della Brexit. In seguito invece riportiamo i nomi delle entità più citate divise per classi. Le classi riportate sono Location, Person, Organisation che sono le 3 classi principali per il task di NER con anche la classe GPE (geo-political entity) in quanto una delle più popolate da questo corpus.

Person

	name	count
0	Brexit	66695
1	Trump	8081
2	Donald Trump	5702
3	Theresa	2628
4	Mrs May	2187
5	Jeremy Corbyn	1800
6	David Cameron	1463
7	Nigel Farage	1427
8	Putin	1411
9	Westminster	1351
10	David Davis	1336

Location

	name	count
0	Europe	17975
1	Asia	1082
2	Africa	818
3	the Middle East	813
4	West	615
5	North	465
6	Atlantic	407
7	North America	359
8	Earth	322
9	Gulf	259
10	Latin America	255

Figure 5: Location e Person

GPE

	name	count
0	UK	58407
1	Britain	34460
2	US	14359
3	London	13789
4	Scotland	13020
5	Theresa	7006
6	U.S.	6946
7	Germany	6194
8	France	5990
9	China	5943
10	Russia	5304

ORG

	name	count
17	European Union	1253
18	the House of Commons	1233
19	State	1202
20	Board	1159
21	AFP	1154
22	the House of Lords	1148
23	Sturgeon	1142
24	Treasury	1122
25	Fed	1111
26	FTSE	1052
27	the European Parliament	1031

Figure 6: GPE e Organisation

Salta subito all'occhio la classificazione dell'entità Brexit in PERSON che si può considerare un errore poichè Brexit è un movimento creato come descritto prima dall'unione di due parole e dovrebbe essere classificata come Organisation. Inoltre è da notare come Trump sia una delle persone più citate in questi articoli sopra gli esponenti politici più importanti del Regno Unito, anche perchè l'entità Theresa May è stata classificata anche come GPE. Tra le altre classi in risalto vi sono la Camera dei Lord e la Cmera dei

comuni con i parlamentari, i quali spesso citati dagli articoli per i continui dibattiti e discussione che si sono verificati ne periodo preso in esame dal dataset fornito.

L'altra libreria utilizzata e che abbiamo visto precedentemente per lo scraping, è **newspaper3k**, la quale forniva la possibilità di estrarre le *keywords* all'interno dell'articolo attraverso il suo metodo di *nlp*(natural language processing).

Figure 7: WordCloud Keywords

Abbiamo sfruttato WordCloud per avere un impatto più semplice visivamente. Come precedentemente visto, le keyword sono molti simili a quelle estratte precedentemente, in particolare la parola più presente, come ci aspettavamo è *brexit*, è seguita dalle entità più coinvolte in questo evento, ovvero *uk* , *eu* e infine *european*. Anche in questo caso la figura di Trump è molto citata all'interno degli articoli, questo è anche dovuto al fatto ad un accostamento si similarità sui comportamenti simili agli Stati Uniti che il Regno Unito ha assunto.

Inoltre vengono mostrati quali sono gli argomenti che fanno discutere di più,il primo in assoluto è *trade*, ovvero il nuovo commercio che dovrà avere il Regno Unito nei confronti di tutto il mondo, in particolare dell'Europa, dopo la loro uscita da essa.

6.2 URL

Un altro aspetto che abbiamo voluto esplorare sono state le fonti dei nostri articoli esaminati. Ogni URL è stato parsato dalla libreria URLLIB per estrarre il dominio e pulirlo dalle stringhe di query e dal protocollo iniziale(http o https).

Una volta estratti i domini sono stati raggruppati per dominio in un nuovo Dataframe, inoltre è stato aggiunto il campo che indica quanti articoli sono stati presi dallo stesso sito online. In questo modo è possibile vedere quali sono i siti che hanno contribuito di più a popolare i nostri dati.

Da questo Dataframe abbiamo potuto costruire dei grafici per mostrare meglio quali siano i siti che hanno influito maggiormente

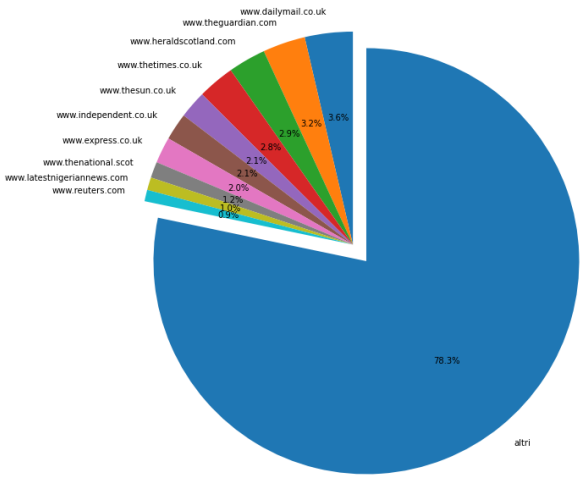

Figure 8: Urls

Come si può notare, dal grafico, nel dataframe quasi 2/3 degli articoli contribuiscono con al più 2 articoli, per questo la percentuale degli "altri" articoli è così elevata.

Per migliorare la visibilità abbiamo deciso di concentrarci solo sui 10 siti più popolosi all'interno del dataset per costruire questo nuovo grafico.

Una volta finito, abbiamo deciso di approfondire ulteriormente la nostra ricerca, eravamo interessati a vedere quali erano i Paesi associati ai domini trovati dai link, per far questo è stato richiesto il contributo di altre 2 librerie: *socket* e *geolite2*. Per associare ogni dominio ad un determinato Paese, la libreria geolite2 si basa sugli indirizzi IP dei server in cui risiedono le notizie analizzate, così al dataframe creato precendemente sono stati aggiunti ulteriori 2 campi: il paese i origine e l'indirizzo IP.

Questo però ha portato dei problemi, in quanto molti siti potrebbero essere associati ad un determinato Paese ma avere un indirizzo IP riconducibile ad un altro paese. Come si vede nell'immagine soprastante,prima di risol-

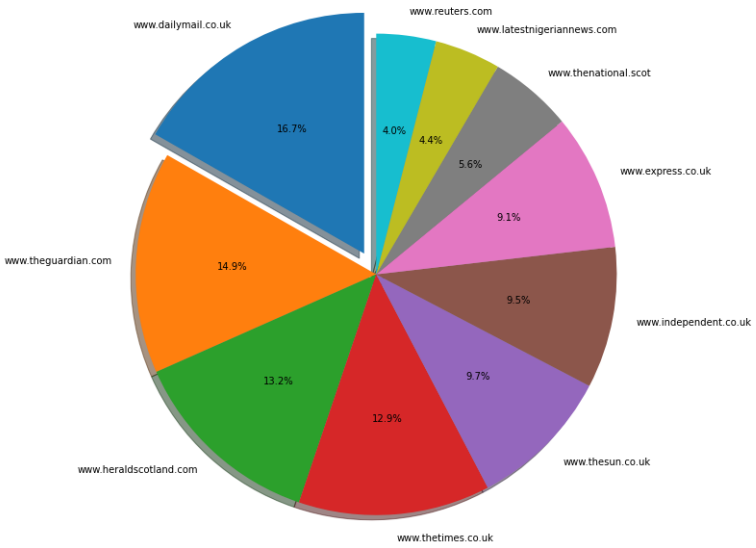

Figure 9: Urls most common

vere il problema, il secondo sito più presente nel nostro dataset, è in realtà un giornale storico del Regno Unito con un indirizzo IP associato agli Stati Uniti.

Per risolvere questo problema abbiamo dovuto controllare a mano le informazioni fornite; essendo una lista lunga, poco più di 4500 diversi siti (4597), abbiamo scelto di controllare solo i giornali che hanno contribuito di più a popolare il DataFrame,con più di 50 articoli. Sempre nella stessa selezione degli domini più comuni, abbiamo scelto anche di indicare quali fossero quei giornali inglesi,appartenenti al Regno Unito, che hanno uno schieramento/allineamento politico che seguono e in particolare se supportano uno specifico partito.

Anche questa analisi è stata fatta ricercando informazioni da fonti sicure, aggiungendo i campi **orientation** per indicare la linea politica del giornale(Destra, Centro-Destra,Centro-Sinistra, etc...). Inoltre alcuni giornali supportano specifici partiti, in questi casi è stato aggiunto un ulteriore campo **party** in cui viene specificato il partito politico che viene appoggiato dalla casa editrice.

Abbiamo scelto di fare questa analisi solo sui giornali del Regno Unito per vedere se determinati giornali appartenente ad un idea politica influenzava o

	url	occorrences	origin	ip	orientation	party
0	www.dailymail.co.uk	983	GB	23.54.196.207	Right-wing	Conservative Party
1	www.theguardian.com	874	GB	151.101.1.111	Centre-left	Labour Party
2	www.heraldscotland.com	772	GB	93.174.10.103	Centrist	NaN
3	www.thetimes.co.uk	756	GB	13.226.50.69	Centre-right	Conservative Party
4	www.thesun.co.uk	567	GB	65.8.183.118	Right-wing	Conservative Party
5	www.express.co.uk	536	GB	65.8.248.71	Right-wing	Conservative Party
6	www.thenational.scot	329	GB	93.174.10.103	Center-left	Scottish National Party
7	www.telegraph.co.uk	222	GB	104.121.97.109	Right-wing	Conservative Party
8	www.irishexaminer.com	195	IE	213.182.13.37	Centrist	NaN
9	www.independent.ie	189	IE	65.8.183.76	Centre-right	NaN
10	www.scotsman.com	172	GB	104.18.2.43	Right wing	NaN
11	www.belfasttelegraph.co.uk	138	GB	13.226.50.43	Centrism	NaN
12	www.mirror.co.uk	124	GB	65.8.183.45	Centre-left	Labour Party
13	www.themalaymailonline.com	122	MY	110.74.168.99	Moderate	NaN
14	www.ft.com	72	GB	151.101.2.109	Centrism	NaN
15	www.newstatesman.com	61	GB	151.101.2.133	Centre-left	Labour Party

Figure 10: Urls con un orientamento politico

meno i propri articoli sulla Brexit. Sono stati trovati 5 orientamenti politici: Destra, Centro-Destra, Centro-Sinistra, Centro e Moderati.

Per i partiti politici invece ne sono stati trovati solo 3: il **Conservative party**, un partito di Destra; gli altri 2 appartengono a partiti di Centro-Sinistra e in particolare sono il **Labour Party** e il **Scottish National Party**. Quest'ultimo,in particolare è un partito politico scozzese.

7 Sentiment Analysis

Per la **sentiment analysis** abbiamo scelto di confrontare diverse tecniche e analizzare i risultati per vedere quale funzionasse di più.

In particolare abbiamo scelto di utilizzare una tecnica classica della sentiment analysis bastata sul *testo*, utilizzando la libreria **Afinn** messa a disposizione per Python e la libreria **Vader**, presente in NLTK che si basa anche lui sul testo ma utilizza una metodolia differente. L'altra tecnica utilizzata è la sentiment analysis basata sugli *aspetti*, conosciuta con l'acronimo **ABSA** (Aspect Based Sentiment Analysis), dopo diverse prove si è scelto di utilizzare la libreria **TextBlob**

7.1 Afinn

Come detto precedentemente, Affin è una tecnica che si basa sul testo, il suo lessico è un elenco di termini inglesi classificati manualmente per valenza con un numero intero assegno compreso tra -5 (negativo) e +5 (positivo).

Come supporto è stata utilizzata la libreria **NLTK**, in particolare per tokenizzare gli articoli. Inizialmente gli articoli sono stati divisi in *sentence*, così da avere una sentiment sulla singola frase dell'articolo, ogni sentence viene trasformata con i caratteri tutti minuscoli. Per dividere gli articoli in sentences è stato utilizzato il modulo *sent_tokenize* presente nella libreria NLTK.

Diviso gli articoli in frasi abbiamo iniziato a lavorare sulle singole sentences, ad ognuna abbiamo applicato lo stesso procedimento: applichiamo il modulo *word_tokenize* per creare una lista formata dalle parole della sentence, successivamente viene applicato il modulo *pos_tag* per tokenizzare e associare ad ogni parola una categoria; questo viene fatto perché per ogni fare vogliamo tenere conto delle entità presenti nella frase, infatti, a parte vengono salvati i "nomi", cioè quelle parole che sono state *taggate* con la sigla NN.(operazione utilizzata per la NER descritta precedentemente)

Infine viene calcolata la sentence su l'intera frase, utilizzando il metodo *score()* dato da Afinn. Le informazioni ricavate vengono salvate in un DataFrame, creato precedentemente, con le colonne: *txt*, *entity*, *sentiment*, in cui ad ogni colonna è associata l'informazione relativa ricavata.

7.2 Vader

Vader è un analizzatore del sentimento fornito dalla libreria NLTK, esso utilizza un elenco di caratteristiche lessicali che sono etichettate come positive o negative in base al loro orientamento semantico per calcolare il sentimento del testo. Il sentimento di Vader restituisce la probabilità che una data frase di input sia positiva, negativa e neutra.

In questo caso, a differenza di afinn, è stato analizzato l'intero articolo, non è stato diviso in sentences, come detto prima il modulo *SentimentIntensityAnalyzer* restituisce un dizionario,composto da 4 elementi, il più importante tra tutti è il campo *compound* esso rappresenta la normale degli altri 3 valori: negativo , positivo e neutro. Se il valore è pari a 0 significa che siamo in presenza di un sentiment neutro, se il valore è positivo rappresenta un sentiment positivo, infine se ha un valore negativo significa che abbiamo una sentiment negativa.

8 Aspect-Base Sentiment Analysis

8.1 Aspect Extraction

Dopo gli step di preprocessing del testo gli articoli sono pronti per l'aspect base sentiment analysis che si dividerà in 2 task: aspect extraction qua descritto e la sentiment analysis che verrà descritta successivamente. Per effettuare l'aspect extraction abbiamo scelto di sfruttare il dependency parser di **Spacy** analizzando ogni frase di un singolo articolo per vedere quali relazioni ci sono all'interno della frase così da estrarre quelli che saranno gli aspetti della nostra analisi insieme agli aggettivi che li descrivono su cui poi verrà fatta la sentiment analysis.

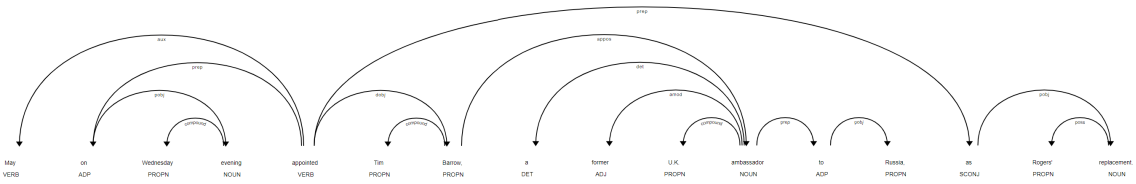

Figure 11: Spacy Reder

Per farlo abbiamo applicato il POS tagging di spacy e abbiamo quindi implementato un approccio rule-based per l'estrazione degli aspetti con la loro descrizione definendo 8 regole: A = aspetto, M = descrizione dell'aspetto

- **Regola 1**: La prima regola è definita così: estrae il token M come descrizione dell'aspetto A se M è in relazione con A come *amod* (Adjective as Modifier)

- **Regola 2**: Assunzione: un verbo avrà solo un *nsubj* e *dobj* Estrai A e M se A è figlio di qualcosa in relazione con *nsubj* , mentre M è figlio di qualcosa in relazione di *dobj*

- **Regola 3**: Adjectival Complement - Assunzione: un verbo avrà solo un *nsubj* e *dobj* A è figlio di qualcosa in relazione di *nsubj* mentre M è figlio dello stesso qualcosa in relazione di *acomp* "The sound of the speakers would be better. The sound of the speakers could be better" - handled using AUX dependency"

- **Regola 4**: Adverbial modifier to a passive verb - Assunzione: un verbo avrà solo un *nsubj* e *dobj* A è figlio di qualcosa in relazione di *nsubjpass*, mentre M è figlio dello stesso qualcosa in relazione di *advmod*

- **Regola 5**: Assunzione: un verbo avrà solo un *nsubj* e *dobj* Complemento di un copular verb, A è figlio di M con relazione di *nsubj* mentre M ha un fglio con relazione di *cop*

- **Regola 6**: Esempio: "It ok", "ok" è un INTJ (interjections come great etc)

- **Regola 7**: ATTR - link between a verb like 'be/seem/appear' and its complement Attributo link tra un verbo come be/seem/appear e il suo complemento Example: 'this is garbage' -¿ (this, garbage)

- **Regola 8**: Per ogni coppia di child controlla dove "A" figlio è l'head di qualche entità e se è vero usa l'intera entità come oggetto Esempio: Air France is cool = [('Air France', 'cool')] invece che [('France', 'cool')]

  Applicando le regole a tutti gli articoli scaricati abbiamo estratto 1mln e mezzo di aspetti con la propria descrizone alla quale è stata applicata la polarità di **Vader**. Alcuni di queste descrizioni sono risultate

neutrali poichè questo approccio tende a raccogliere anche gli aggettivi qualificativi che non hanno una polarità in quanto non aggravano e non migliorano la qualità del sostantivo a cui si riferiscono. Questa wordcloud rappresenta la frequenza degli aseptti estratti

Figure 12: WordCloud aspetti

In questo insieme di articoli riguardanti la Brexit si parla di governo, economia, voto, referendum, mercato ecc.. che sono gli aspetti principali di cui si è parlato tra politici, stampa internazionale e nazionale e sono stati oggetti di discussione per lungo tempo. Gli aspetti estratti sono molti e il passo successivo prevede di provare a clusterizzare gli aspetti con l' algoritmo **Kmeans**.

8.2 Aspect Clustering

Per la clusterizzazione degli aspetti è stato implementato k-means con un numero variabile di cluster da 2 a 10, ad ogni risultato è stata applicata l'analisi di silhouette per selezionare il numero di cluster ottimale.

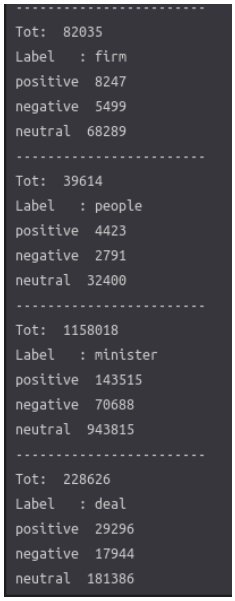

Figure 13: Clusters

IL miglior risultato è stato trovato con 4 cluster seppure la media del punteggio di silhouette è di poco superiore a 1. I cluster trovati sono : **deal**, **people**, **minister**, **firm**. Il risultato non è stato molto soddisfacente, infatti essendo gli articoli estratti per il topic Brexit essi e gli aspetti estratti hanno molta similarità tra loro. Anche in termini di Sentiment Analysis i cluster erano troppo generali e la maggior parte degli articoli è di polartà positiva se preso al completo, di conseguenza anche i cluster risultano di polarità positiva. I risultati della polarità di Vader per i quattro cluster sono mostrati nelle sezioni successive. Gli stessi cluster però contengono altri sottoinsiemi di aspetti come trade-market-economy oppure government-minister- vote-referendum-poll. Esamindando i cluster sono stati raccolti altri insiemi di aspetti più specifici sui quali abbiamo basato i risultati dell' Aspect-Base Sentiment Analysis. In questo insieme di articoli riguardanti la Brexit si parla di governo, economia, voto, referendum, mercato che sono gli aspetti principali di cui si è parlato tra politici stampa internazionale e nazionale e sono stati oggetti di discussione per lungo tempo. Gli aspetti estratti sono molti il passo successivo prevede di provare a clusterizzare gli aspetti con l' algoritmo **Kmeans**.

8.3 Aspect Base Sentiment Analysis

Esaminando i cluster creati da *k-means* abbiamo raccolto gli aspetti più frequenti che sono rappresentati da token affini e che rappresentano la stessa entità o che formano un topic specifico. In questa analisi più approfondita abbiamo estratto diversi topic ognuno dei quali raccoglie gli aspetti più frequenti associati ad esso basandoci anche sulla frequenza generale degli aspetti. Prima di estrarli quindi ci siamo chiesti quali fossero gli aspetti che caraterizzano questo evento, appunto esaminando la wordcloud delle frequenze degli aspetti ci è risultato che i topic più discussi sono: **Economia** sul lato economico si esprimono opinioni e preoccupazioni sul futuro anche in base agli **Accordi** (altro topic) commerciali ed economici che si trattano con l'Unione Europea e grande oggetto di discussione , abbiamo quaindi estratto il topic **Europa** o unione europea per capire l'opinioni dei giornali sulla **Brexit** anch'essa estratta come topic. Un altro topic estratto è il **Referendum** su cui ci sono state molte controversie sul voto e sul risultato discutendo il fatto di lasciare al popolo britannico la decisione sull'uscita dall'Europa referendum vinto dal SI per pochi voti 52% vs 48%, UN altro tema estratto è l' **Immigrazione**, oggetto tanto discusso come conseguenza della Brexit poichè non basta avere la cittadinanza europea per raggiungere il Regno Unito e soprattutto preoccupava la situazione degli immigrati che in quel periodo lavoravano in quei paesi, tanta incertezza attorno a questo topic, infine abbiamo preso in considerazione gli U.S.A come topic, essendo Trump una delle entità più citate, insieme al fatto che gli Stati Uniti così come il resto del mondo hanno avuto un forte interesse e preoccupazione per la Brexit, dovendo ridefinire accordi commerciali ed economici. L'ultimo aspetto-topic preso in esame è la **Politica** che è oggetto e soggetto di quasi tutti gli articoli etsratti poichè i parlamentari, il primo ministro **Theresa May** e il suo partito **Conservative party**, il partito a capo dell'opposizione **Labour Party** e il suo leader **Jeremy Corbyn** sono stati i porincipali protagonisti di questo evento, così come Nigel Farage spesso nominato. In seguito è mostrata una tabella che raccoglie tutti i topic con i loro aspetti, anticipando i risultati della polarità che verranno approfonditi nella sezione successiva.

Topic	Aspetti
Labour party	”corbyn” , ”labour”
Conservatve party	”conservative”, ”theresa”, ”may”
Europe	”brussels”, ”europe”, ”union”, ”europeans”, ”eurosceptics”, ”germany”, ”italy”, ”euro”, ”eu”
Brexit	”brexit”
Referendum	”referendum”, ”vote”, ”election”, ”brexit”, ”poll”, ”exit”
Deal	”deal” , ”firm”, ”meeting”, ”offer”, ”future”, ”negotiation”,”decision”, ”bill”
Economy	”trade”, ”economy”, ”market”, ”bank”, ”industry”, ”price”, ”pound”, ”tax”, ”money”, ”economist”
Immigration	”migration”, ”immigrant, ”migrant”, ”muslim”, ”immigration”
USA	”us”, ”u.s”, ”trump”, ”dollar”, ”donald”
UK	”uk”, ”country”, ”britain”, ”kingdom”, ”london”, ”scotland”, ”england”

9 Risultati

9.1 Aspect Based Sentiment Analysis

9.1.1 Risulati Clustering

I risultati del clustering sono tutti positivi poichè troppo generali, e come visto nell'analisi della sentiment generale sugli articoli la polarità risulta preponderante verso la positività, in seguito si riportano i risultati delle polarità divisi per clustering.

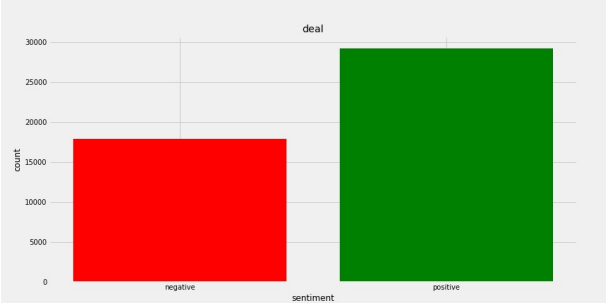

Figure 14: Deal

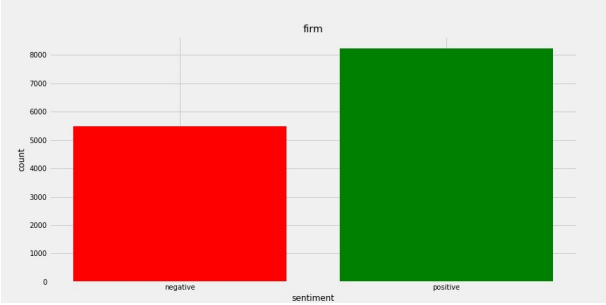

Figure 15: Firm

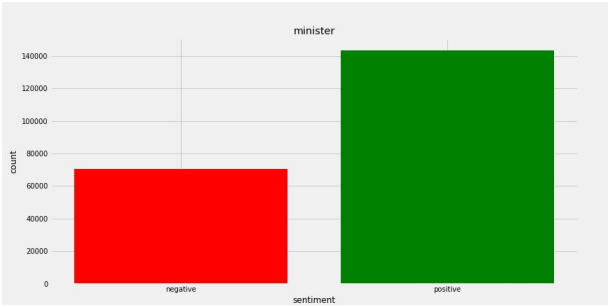

Figure 16: minister

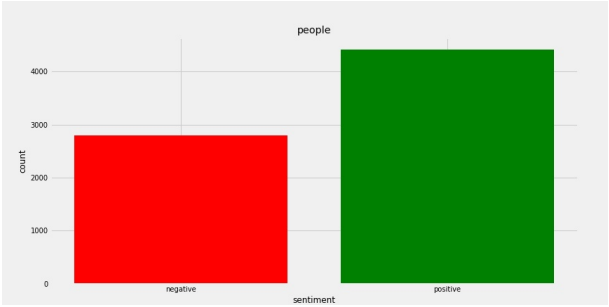

Figure 17: people

I risultati di questi cluster generali sono tutti positivi in quanto seguono la distribuzione di polarità generale degli articoli che come si vede dai risultati è per la maggior parte positiva. A questo livello di granularità degli aspetti non si riescono a trovare degli insights comprensibili e validi.

9.1.2 Cluster con granulatura maggiore

Come detto in precedenza i risultati del clustering sono troppo generali, il che porta ad avere dei risultati non proporzionati a favore degli aspetti con una sentiment positiva.

Per migliorare i risultati abbiamo deciso di creare dei cluster con una granularità maggiore, questo avviene perché siamo andati noi a selezionare gli aspetti di ogni cluster con un affinità maggiore tra di loro.

Dopo aver selezionato i nuovi cluster di aspetti,creando cosi dei dei topic, abbiamo deciso di confrontare la sentiment di ogni giornale al quale abbiamo

associato un orientamento politico. In questo modo speriamo di andare a notare quanto un giornale possa influenzare in merito agli argomenti trattati.

Per far ciò sono stati presi quegli articoli di un determinato giornale e gli aspetti che rientravano nel topic da analizzare, veniva presa la sua sentiment, ovvero il valore di *compound* e fatta la media.

9.1.3 USA

Polarità

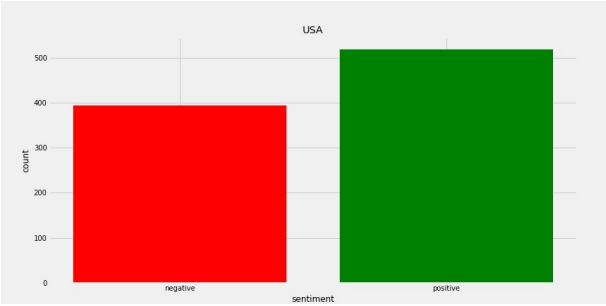

Figure 18

Wordcloud aggettivi positivi

Figure 19

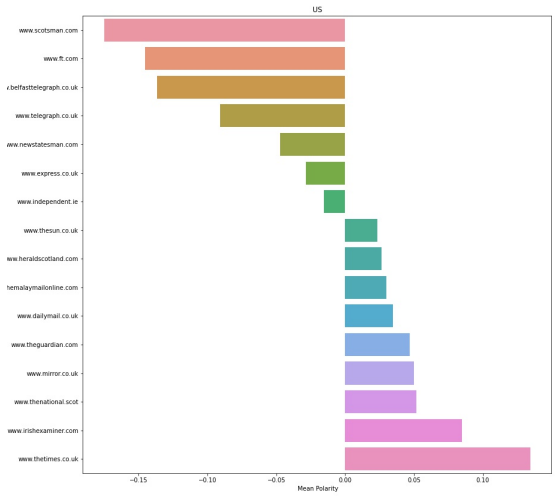

Figure 20

Questo topic è interessante perché il Regno Unito è stato accostato molto agli Stati Uniti per la nuova indipendenza che otterrebbe una volta conclusa la Brexit. La polarità su questo topic è positiva e dalla wordcloud si può notare come gli stati uniti vengono definiti dai giornali come **Important**, **powerful**, **strong**, **defensive**, **isolationist** tutti aggettivi che caratterizzano la superpotenza mondiale come un'importante pedina nel processo della Brexit come ruolo esterno per accordi commerciali ed economici. Da quello che mostra il grafico, non ci sono particolari correlazioni tra i partiti, infatti i giornali che presentano gli aspetti più negativi e positiva riguardanti questo cluster sono associati entrambi a partiti di destra, in particolare rispettivamente ad un partito di Destra e uno di Centro-Destra.

9.1.4 UK

Polarità

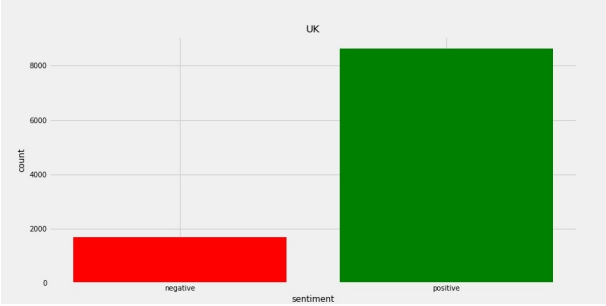

Figure 21

La polarità è estremamente positiva, la maggiorparte degli articoli e dei giornali proviene dal Regno Unito e nonostante il controverso periodo gli articoli mantengono una certa appartenenza agli stati che compongono il Regno.

Wordcloud aggettivi positivi

Figure 22

Anche in questo caso sono stati presi in esame gli aggettivi che contribuiscono alla polarità vincente , quella positiva e come si può notare gli aggettivi che più contribuiscono sono united, important , best good great e strong, aggettivi visti anche nel topic USA, da notare come questo aspetto

viene definito anche come *rich* e *free* ad indicare come i giornali descrivono il Regno Unito nell'immediato futuro.

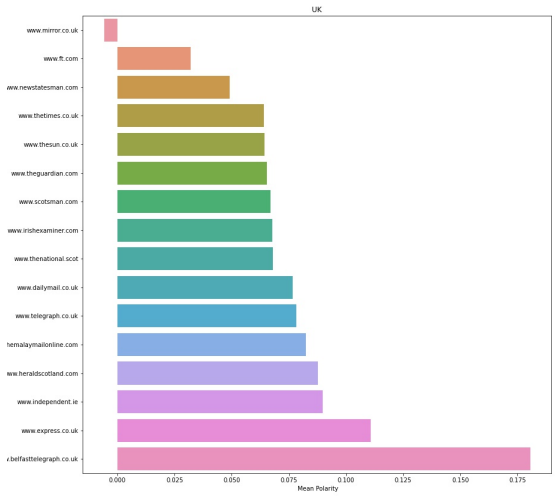

Figure 23: polarità media giornali

In questo caso ci troviamo in un Topic i cui aspetti sono sempre positivi tranne che per un giornale, il *Mirror*, esso è un giornale associato ad un orientamento di sinistra il quale ha supportato più volte il Labour party.

9.1.5 Economy

Polarità

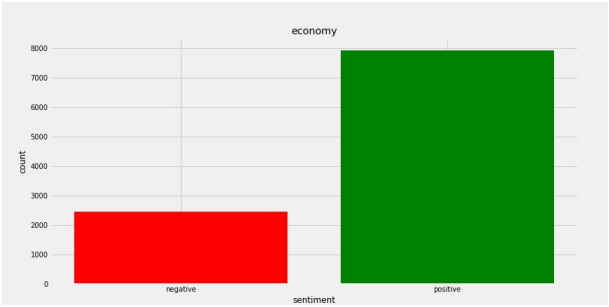

Figure 24

.

Wordcloud aggettivi negativi

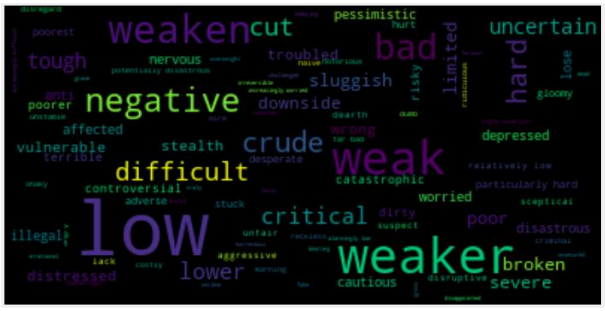

Figure 25

Wordcloud aggettivi positivi

Figure 26

L'aspetto economico è stato uno degli aspetti più discussi in questo corpus di articoli mettendo in risalto il fatto che secondo queste analisi la prospettiva che risulta da questo aspetto è molto positiva. Andando a guardare gli aggettivi utilizzati quelli negativi risaltano la debolezza economica che molto probabilmente è riferita più all'economia europea che influisce su quella nazionale, d'altra parte si vede come l'aggettivo più usato in questo topic è **free** ad indicare come uscendo dall'europa si avrebbe un'economia e un mercato più libero meno soggetto a restrizioni imposte dall'UE.

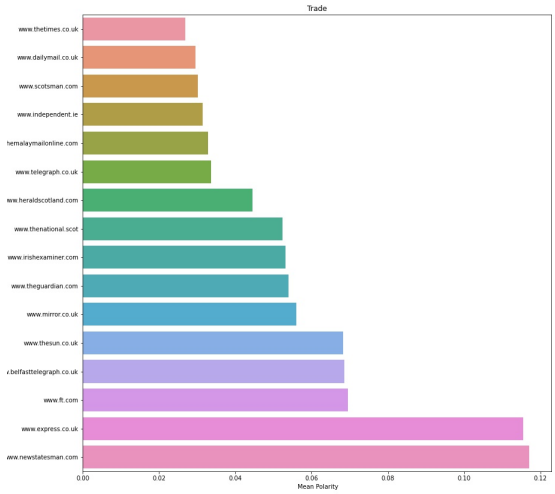

Figure 27

Questo cluster ha una sentiment molto simile al topic visto precedentemente, anche se in questo caso sono tutti positivi.

Questo topic, al contrario degli altri. riguarda una persona, Theresa May, il Primo Ministro che ha dato inizio alla Brexit

9.1.6 Theresa May- Conservative party

Polarità

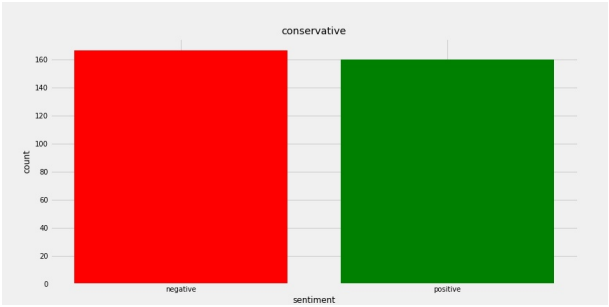

Figure 28

Wordcloud aggettivi negativi

Figure 29

Wordcloud aggettivi positivi

Figure 30

Seppur di poco la polarità associata a Theresa May e al suo partito il Conservative party è negativa, Theresa May è sicuramente una figura controversa, da primo ministro ha cercato in tutti i modi di concludere la Brexit e quindi di stringere accordi con l'UE per l'uscita del Regno Unito ripetutamente rifiutati dal parlamento britaanico. Anche il suo partito è diviso in due sulla sua figura e sulla Brexit da lei tanto "spinta". Gli articoli presi in esame infatti con gli aggettivi negativi non si trattengono, definita *shameful*, *rebel*, *disappointed*, *not clear*.

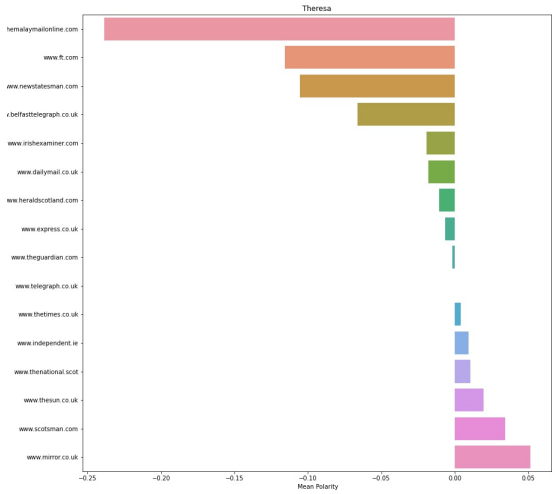

Figure 31

Theresa May era il Primo Ministro e leader del partito Conservatore, un partito di Destra, infatti come notiamo dal grafico i giornali che hanno una sentiment negativa sono quelli che supportano partiti di Sinistra o del Centro. In particolare il *the malay mail online* è un giornale appartenente alla Malesia e ha un allineamento polito con i Moderati.

Da notare che sono presenti anche un paio di giornali che sostengono una filosofia di Destra, questo può essere dato dal fatto che durante queste trattative, il partito di Theresa May non era tutto unito, vi erano pareri contrastanti all'interno dello stesso partito.

9.1.7 Labour party -Jeremy Corbyn

Polarità

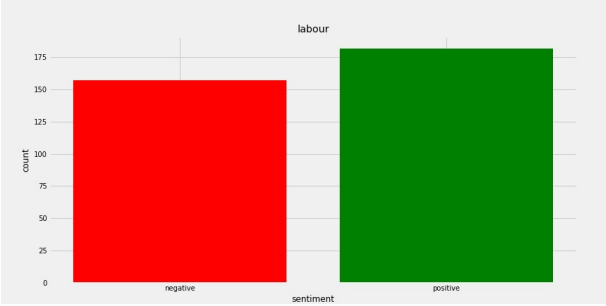

Figure 32

Wordcloud aggettivi positivi

Figure 33

Il Labour party era a capo dell'opposizione in parlamento nel periodo di copertura del dataset, come ci aspettavamo se la polarità sul Conservative Party e THeresa May è negativa, la polarità riguardante l'opposiziuone politica è positiva. Questa è una dinamica che avviene anche in Italia dove chi è gestisce il governo è sempre criticato molto di più rispetto all'opposizione che si occupa di trovare i problemui e di criticare a sua volta l'operato del governo.

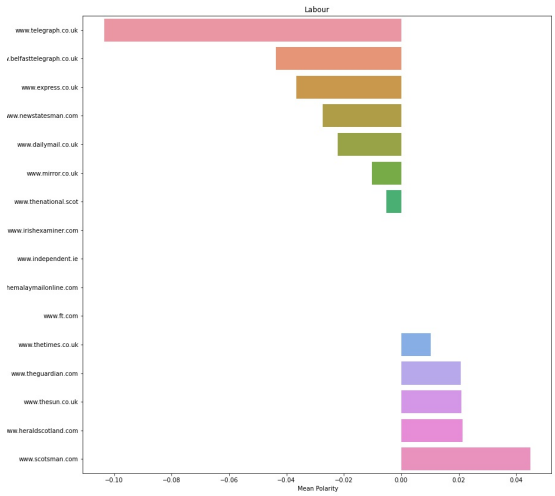

Figure 34

Questo topic mostra comportamenti molto simili al cluster di Theresa, come in quel caso, le sentiment più negative sono da parte di quei giornali che hanno un allineamento di Destra. Mentre la maggior parte di quelli di Sinistra riportano dei sentiment positivi.

In questo caso, in più degli altri, vi sono 3 giornali che non hanno una sentiment in quanto negli articoli presenti nel DataFrame non sono stati citati gli aspetti descritti precedentemente

9.1.8 Referendum

Polarità

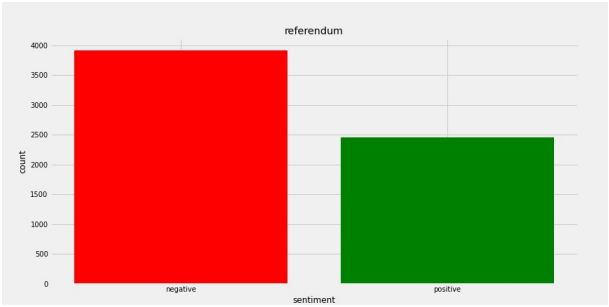

Figure 35

Wordcloud aggettivi negativi

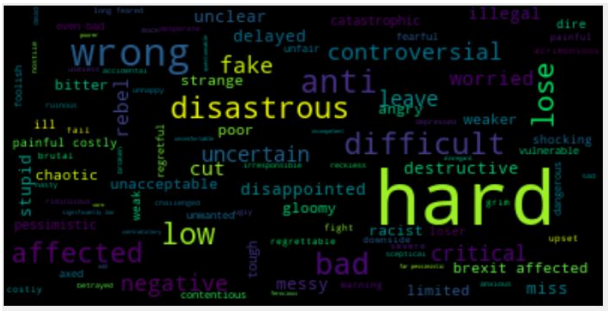

Figure 36

Wordcloud aggettivi positivi

Figure 37

IL referendum è stato vinto con quasi 52% dei voti per lasciare l'Europa, è stato molto discusso e definito come *fake*, *controversial*, *wrong*, *disastrous* che come si vede dall'immagine seguente sono tutti aggettivi usati da giornali che sono contro la Brexit e quindi anche contro il risultato del referendum, d'altra parte gli aggettivi positivi più utilizzati in questo topic sono *popular*, *meaningful*, *fresh* e *important*, anche qui i giornali con orientamento politico più europeista si sono schierati a favore.

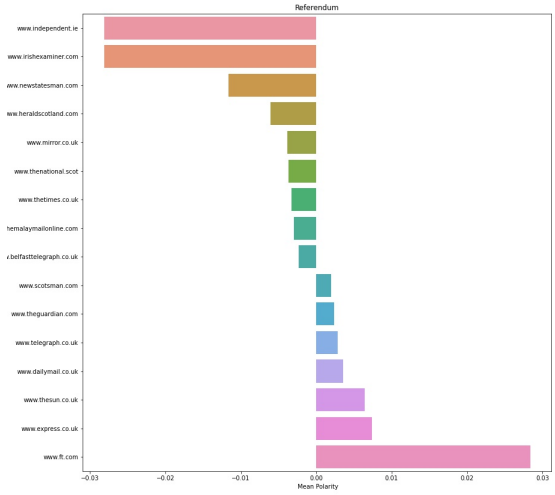

Figure 38

Con questo cluster abbiamo una vera divisione su quali siano i giornali che spingono e supportano la Brexit e quelli che invece vorrebbero rimanere con l'Europa. Infatti i partiti di Sinistra presentano una sentiment negativa sul Referendum, al contrario di quelli di Destra che mostrano una sentiment positiva.

9.1.9 Europe

Polarità

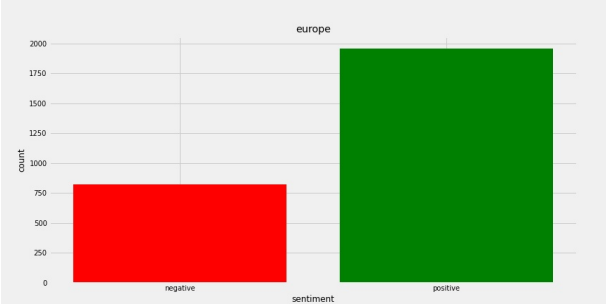

Figure 39

Wordcloud aggettivi negativi

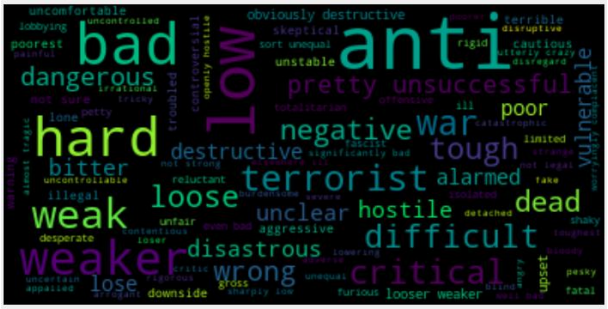

Figure 40

Wordcloud aggettivi positivi

Figure 41

La polarità espressa sull'europa è sorprendentemente positiva, questo vuol dire che quando andremo ad analizzare il topic Brexit ci aspettiamo una polarità opposta. Gli articoli presi in esame quindi esprimono la contrarietà sull'uscita dall'europa del Regno Unito descrivendo l'europa come *strong*, *ambitious secure* ecc.. Mnetre chi esprime opinioni negative descrive l'europa soprattutto come *weak dangerous* e addirittura *terrorist*.

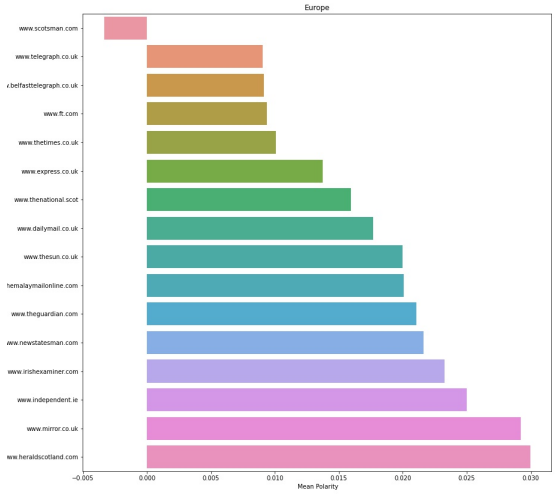

Figure 42

Anche in questo cluster, abbiamo quasi tutti i valori positivi, indistintamente dall'orientamento politico dei giornali. Solo un giornale di Centro-Destra ha una sentiment negativa, il *The Scotsman*, probabilmente per screditare l'Europa e sostenere la Brexit.

9.1.10 Brexit

Polarità

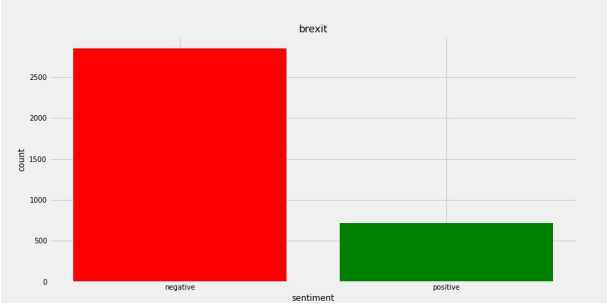

Figure 43

Wordcloud aggettivi negativi

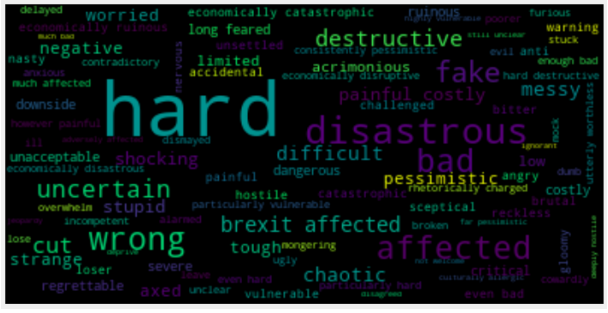

Figure 44

Wordcloud aggettivi positivi

Figure 45

Se la polarità riferita agli aspetti del topic Europa sono risultati positivi, per l'aspetto Brexit la polarità è negativa, questo ha senso in quanto la maggiorparte dei giornali sembra aver espresso quindi perplessità sulla brexit e le conseguenze che questa avrà sul Regno Unito, gli aggettivi negativi più utilizzati sono *disastrous*, *bad*, *uncertain* a sottolineare l'incertezza di questa decisione drastica. In particolare un aggettivo usato molto spesso è *hard*, negli articoli analizzati è spesso usato il termine hard-Brexit in riferimento al fatto che gli accordi con l'UE sono difficilmente applicabili e come si vede nel grafico sottostante nel mese di Gennao 2017 c'è un picco di negatività causato dall'utilizzo dell'aggettivo hard poichè in quel mese Theresa May annunciava che la Brexit si farà con o senza accordi con l' UE.

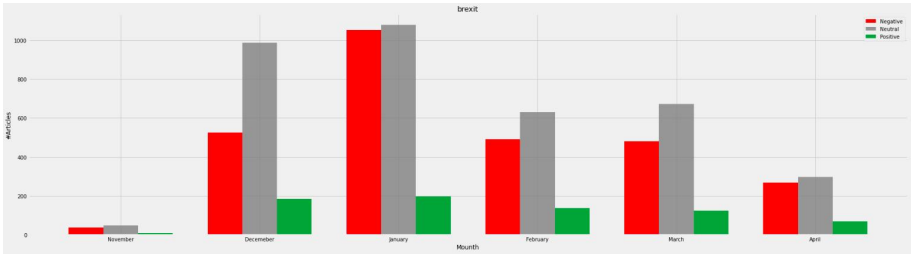

Figure 46

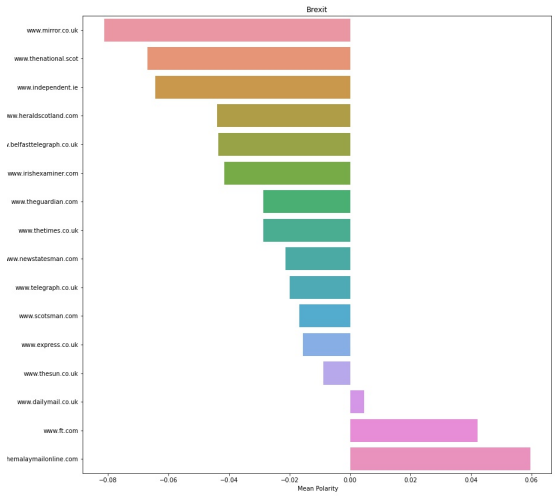

Figure 47

Per questo caso particolare, abbiamo voluto analizzare solo l'aspetto brexit, da come si può vedere dal grafico tutti i partiti di Sinistra hanno una sentiment negativa, ma anche quei giornali con un orientamento politico di Destra o Centro che non sono inglesi, ma irlandesi o scozzesi come il *HeraldScotland* e il *Independent*

9.1.11 Deal

Polarità

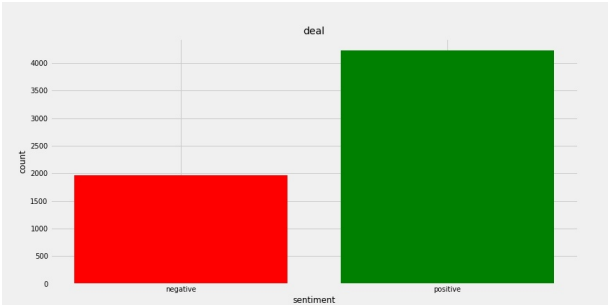

Figure 48

Wordcloud aggettivi negativi

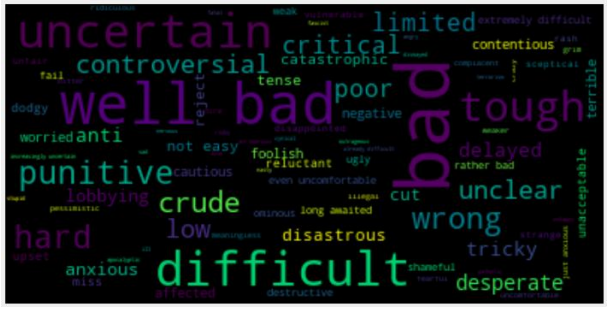

Figure 49

Wordcloud aggettivi positivi

Figure 50

Se prima abbiamo parlato di hard Brexit è perchè gli accordi con l'UE sono molto difficili e l'Unione lascerà uscire il Regno Unito con degli accordi senza sconti. La polarità associata a questo topic è positiva e questo ci indica come ci sia ottimismo per la buona uscita del Regno Unito descrivendo gli accordi come *comprensive*, *great* come si vede nella wordcloud.

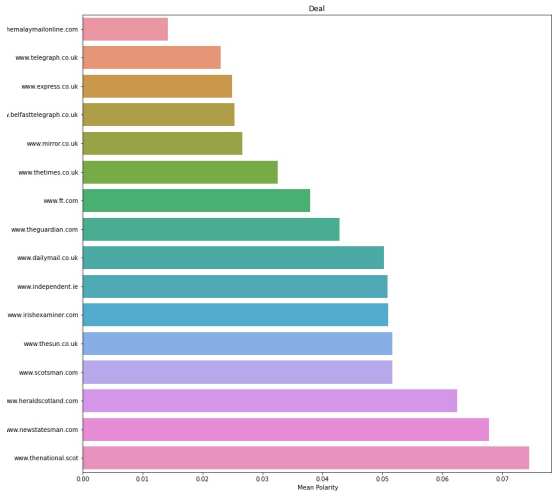

Figure 51

In questo caso abbiamo tutte le sentiment positive, anche qui l'orientamento politico dei giornali non hanno influenzato gli articoli scritti.

9.1.12 Immigration

Polarità

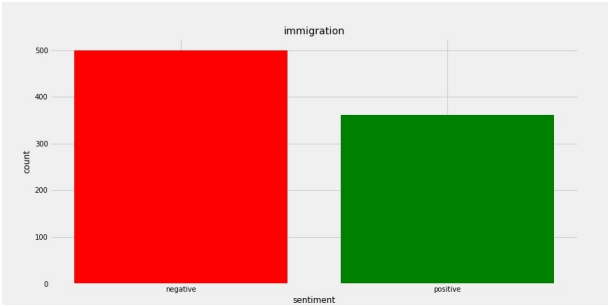

Figure 52

Wordcloud aggettivi negativi

Figure 53

Wordcloud aggettivi positivi

Figure 54

La polarità associata al topic dell'immigrazione è negativa e questo non stupisce. Infatti uno degli aspetti più importanti e su cui la Brexit ha avuto un appoggio importante è quello della gestione dell'immigrazione da parte dell'UE e dei numerosissimi cittadini europei presenti sul suolo britannico. Si è parlato spesso infatti di una gestione interna dell'immigrazione uscendo dall'europa, si parla infatti soprattutto di *illegal uncontrolled*, *terrorist*, tutte opinioni espresse dai giornali e dai politici che hanno avuto una grossa influenza sul risultato della Brexit

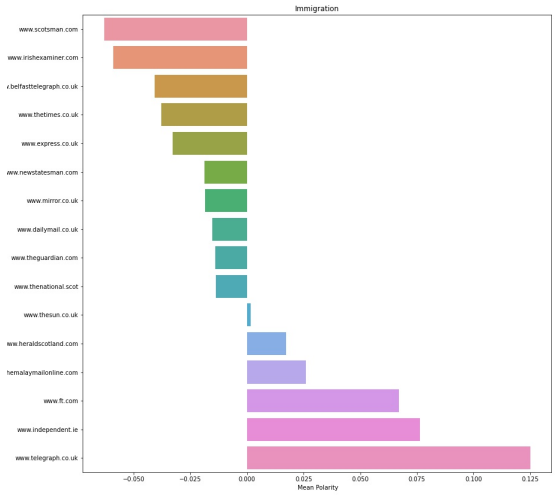

Figure 55

9.2 Text Sentiment Analysis

Come detto precedentemente per la sentiment analysis basata sul testo ci siamo affidati a 2 librerie, *Afinn* e *Vader*. Abbiamo scelto di analizzare solo i risultati avvenuti con Vader, in quanto, per il percorso effettuato e l'utilizzo prevalente della libreria NLTK era la scelta più congrue e affine.

Inizialmente abbiamo analizzato tutti gli articoli presenti nel nostro dataset, per capire in generale quale fosse la sentiment a livello di gruppo

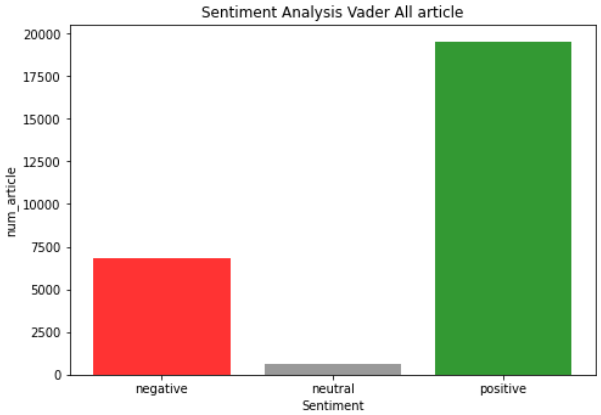

Figure 56: Vader all article

Vediamo come il sentimento positivo sia in maggioranza rispetto a quello negativo, a differenza con afinn, Vader ha meno articoli con una sentiment neutra, inoltre notiamo che il dataset è fortemente sbilanciato, in quanto i negativi sono meno di 1/3 rispetto ai positivi(6851:negativi, 19548: positivi, 609: neutrali)

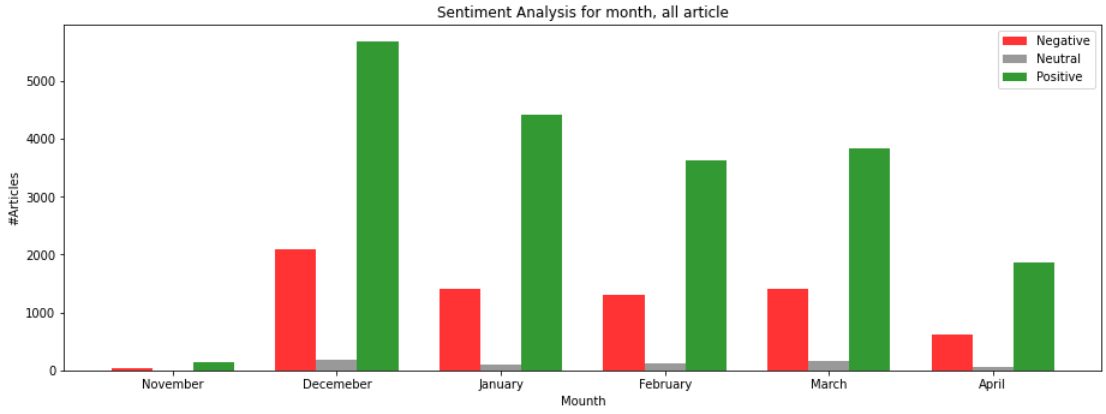

Figure 57: Vader all article for month

Raggruppando gli articoli per mese si nota che la sentiment più alta degli articoli è sempre quella positiva, ma non è costante, a differenza della sentiment negativa, che varia in maniera minore rispetto alla sentiment positiva.

Dopo questa analisi su l'intero dataset abbiamo scelto di selezionare maggiormente il dataset. Ci siamo chiesti qual'era la sentiment degli altri paesi,al di fuori del regno unito, così si è preso tutti gli URL con il campo 'origin' diverso da GB.

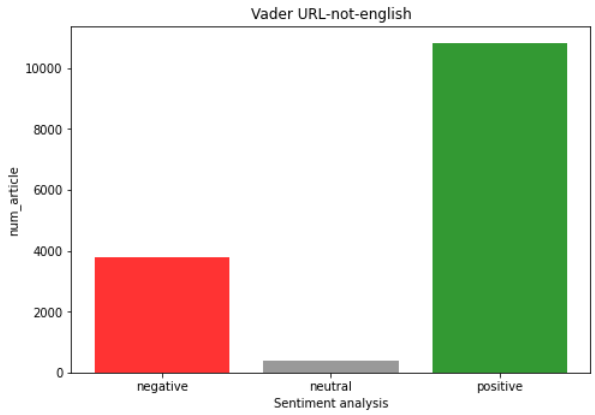

Figure 58: Vader article not english

In questo caso abbiamo un rapporto sempre a favore di una sentiment positiva ma minore rispetto a quella vista con tutto il Dataframe. Adesso gli articoli che hanno una sentiment negativa sono più di un terzo rispetto a quelli con una sentiment positiva.

Topic	Aspetti	Polarità
Labour party	”corbyn” , ”labour”	POSITIVE
Conservatve party	”conservative”, ”theresa”, ”may”	NEGATIVE
Europe	”brussels”, ”europe”, ”union”, ”europeans”, ”eurosceptics”, ”germany”, ”italy”, ”euro”, ”eu”	POSITIVE
Brexit	”brexit”	NEGATIVE
Referendum	”referendum”, ”vote”, ”election”, ”brexit”, ”poll”, ”exit”	POSITIVE
Deal	”deal” , ”firm”, ”meeting”, ”offer”, ”future”, ”negotiation”,”decision”, ”bill”	POSITIVE
Economy	”trade”, ”economy”, ”market”, ”bank”, ”industry”, ”price”, ”pound”, ”tax”, ”money”, ”economist”	POSITIVE
Immigration	”migration”, ”immigrant, ”migrant”, ”muslim”, ”immigration”	NEGATIVE
USA	”us”, ”u.s”, ”trump”, ”dollar”, ”donald”	POSITIVE
UK	”uk”, ”country”, ”britain”, ”kingdom”, ”london”, ”scotland”, ”england”	POSITIVE

10 Conclusioni

In generale quindi come motsrato in questa tabella la maggior parte degli aspetti sono positivi, ad eccezione del topic Brexit, immigrazione e Theresa May. Diciamo che le opinioni espresse in questo corpus di articoli pongono l'attenzione sulle conseguenze che potrebbe avere la Brexit esprimendo opinioni negative al riguardo mentre sull'europa si esprimono opinioni positive. Gli altri aspetti sono stati classificati come positivi poichè molto spesso gli articoli riportano le dichiarazioni fatte da politici, economisti e abbiamo visto che seppur cercando di rimanere obiettivi alcuni giornali tendono a esprimere un'opinione più esplicita verso una o l'altra direzione. Il politico sicuramente più citato e da cui sono riportate le dichiarazione è Theresa May che sotto il punto di vista economico, finanziario e commerciale vedere nella Brexit una rinascita per il proprio paese. Molti aspetti raccolti sono neutrali, non solo perchè l'algoritmo ruled-base ha raccolto aggettivi qualificativi senza un'esplicita polarità, ma anche perchè un giornale ed un giornalista in teoria cercano di riportare l'articolo e i fatti con obiettività senza sblianciarsi in opinioni esplicite come può succedere per le recensioni di un prodotto e se

spesso sono di parte per un determinato argomento esprimono la loro opinione in modo molto implicito. Questo ha introdotto delle difficoltà rilevanti, ma aumentando il livello di granularità degli aspetti questo progetto è riuscire a trovare degli insights validi e comprensibili.

10.1 Sviluppi Futuri

Per migliorare la qualità delle analisi si potrebbe applicare un modello di topic modeling per l'aspect extraction come **LDA** al posto dell'algoritmo *ruled-base*, inoltre si è provato ad applicare **JST** con scarso successo in termini di risultati risorse e tempo. Un altro sviluppo futuro sarebbe quello di applicare **JST** e **ASUM** per confrontare i risultati e gli aspetti trovati.