The New York Times et ses lecteurs — une analyse de sentiments
Machine Learning for Natural Language Processing 2020

Abdulrahman Kassab
ENSAE Paris
abdul.rahman.kassab@ensae.fr

Youssr Youssef
ENSAE Paris
youssr.youssef@ensae.fr

Abstract

Le New York Times est un média grand public ou un *household name* et brasse un grand nombre de lecteurs. Est-il alors toujours au diapason avec son lectorat ? A travers les problématiques abordées, les sentiments qui se cachent derrière les grands titres, et les commentaires des lecteurs, cette analyse se propose de mettre en lumière les déphasages entre les journalistes ou éditorialistes et celui qu'on appelle "le grand public".

1 Problématique

Dans le cadre de notre analyse, visant à analyser l'évolution des sentiments exprimés par le journal et par ses lecteurs - dans le contexte politique de l'élection de Donald Trump - nous nous concentrons sur les articles d'informations (type News) et éditoriaux (type op-Ed). En effet, le NYT étant un journal très exhaustif, nous ne voulons pas polluer notre analyse en mélangeant les articles sur l'actualité politique ou économique américaine avec des articles de cuisine.

Certains thèmes comme la politique, l'environnement... sont-ils plus susceptibles d'être commentés ? Ou est-ce plus lié au type de titre ; un sentiment polarisé attirerait-il des personnes qui, en retour, feraient des commentaires polarisés ? La longueur de l'article joue-t-elle également un rôle ?

2 Protocole expérimental

Données

La *database* du New York Times est composée de données sur des articles publiés de Janvier à Mai 2017 ainsi que de leurs commentaires mais aussi d'articles publiés de Janvier à Avril 2018 ainsi que leurs commentaires.

Nous nous retrouvons donc avec huit tableaux, pour chaque année, une analyse sur les articles d'informations (deux tableaux : un pour les headlines, un autre pour les commentaires) et une autre sur les articles éditoriaux (idem).

Modèles utilisés

Pour mener à bien notre analyse[1], nous avons recours à deux outils majeurs, TextBlob, Word2Vec et Blob.

Dans un premier temps, TextBlob nous permet d'obtenir un indicateur de sentiment concernant chaque titre d'article (headline) et chaque commentaire. Celui-ci nous donne un indicateur de sentiment entre -1 (sentiment négatif) et 1 (sentiment positif). 0 indique un sentiment neutre.

Nous avons procédé à trois analyses grâce à Word2Vec. La première consiste à effectuer des clusterings après avoir train nos données. Ceux-ci sont visibles sur le notebook[2] et permettent de se faire une idée des associations de mots dans les titres et les commentaires ainsi que de la pertinence de nos modèles. Le cas échéant, on retrouve une plus grande logique dans les clusters générés par les commentaires, ce qui s'explique par la taille significativement supérieure des corpus de commentaires. La seconde consiste à comparer les mots les plus proches à un groupe de mots importants au contexte politique : Trump, Republicans, Democrats, Tax et Immigration (ici pour les articles éditoriaux de 2017 par exemple). La dernière évalue la proximité de ces mots à good et bad pour en tirer le sentiment exprimé dans les titres et dans les commentaires (Trump est-il vu positivement ?)

Evaluation des modèles

Nous avons vérifié la pertinence de textblob en faisant apparaître au hasard 5 éléments positifs,

[1] Lien vers le Google Colaboratory de l'étude
[2] Lien vers les cellules du Colab

5 neutres et 5 négatifs. Ainsi, un exemple de commentaire positif dans les headlines est “Dinner, That Beautiful Dance”; un exemple de snippet neutre est “Patients who continued to take the cholesterol-lowering drugs on the day of surgery had a 48 percent reduced risk of dying in the next 30 days” et un exemple de commentaire négatif est: “I dare anyone in the Justice Department to go and look at Breitbart News and other so called Alt-Right websites. Then read the vile and ignorant comments left by their followers. Not to mention the clearly racist imagery they post. Then you’ll see why having Steven Bannon as Chief Strategist in the Donald Trump administration is so disturbing.“

Nous avons également procédé à une évaluation quantitative via le F-score et qualitative via l’étude des mots les plus similaires.

Nous reprenons la fonction d’évaluation du TD3 qui nous donne trois évaluations quantitatives : précisions, recall et f1-score ainsi que la database de tweets sur Virgin America. Nous utilisons notre modèle issus de l’évaluation des commentaires sur les articles d’informations de 2017 afin de prédire des labels sentiments [3].En effdet, c’est cette database qui contient le plus de mots.

3 Résultats

Analyse préliminaire

Une analyse statistique préliminaire montre qu’en 2017, l’écart-type du nombre de commentaires était plus grand pour les News, en 2018 le déséquilibre s’inverse. Il y a plus ou moins 100 commentaires d’écart entre les écart-types. Dans tous les cas, les articles d’opinions ont plus de commentaires que les articles de news pures.

Sur l’année 2017 par exemple, nous voyons que la distribution des sentiments des titres d’articles (voir Figure 1) est autour de 0 alors que celle des commentaires (voir Figure 2) est beaucoup plus étalé. Ce résultat est logique étant donné que le journal est censé être impartial donc objectif et neutre alors que les lecteurs livrent un avis subjectif (donc orienté) en commentaire.

Les wordclouds (voir Figure 3) que nous avons générés montrent également la prédominance du Président Trump et dans les commentaires, et dans les titres d’articles, ce qui montre son omniprésence dans l’actualité et dans l’inconscient collectif.

[3] Lien Cellule Collab

Mais il est aussi intéressant de voir que les problématiques évoquées par les journaux via leur headlines ou snippets, ne sont pas nécessairement les problématiques les plus commentées. On voit par exemple que le biais politique n’a été repris dans les commentaires qu’en 2018 par rapport à Donald Trump. Les commentaires mettent en avant d’autres problématiques moins évoquées comme le système de santé. Il y a-t-il une interaction, si ce n’est thématique, mais ”sentimental” entre les journalistes et éditorialistes et leur lectorat ?

Analyse data science - Word2Vec

[4]

On voit que le mot Trump est associé en 2017 plus négativement dans les commentaires que dans les headlines. De même, immigration semble être plus négatif sur l’analyse des headlines que dans les commentaires. Quand au mot tax , le mot est associé à des mots plus négatifs dans les commentaires que dans les headlines que se soit en 2017 ou en 2018.

On fait aussi apparaître les mots les plus proches selon commentaires/headlines pour chaque année. Par exemple, pour les articles éditoriaux 2017 (voir Figure). Les résultats montrent un écart entre les points de vues exprimés par les édiotiralistes et les lecteurs. Les premiers semblent favorable à l’immigration[5] : bill pour projet de loi, help dans les mots les plus proches) alors que les seconds moins (deportation , anti-immigration). On retrouve ce même écart dans les différentes associations des tableaux good / bad , où les lecteurs semblent moins favorables à l’immigration[6]. Le test de notre modèle basé sur les commentaires des articles d’informations en 2017 donne des résultats satisfaisants. [7]. Nous retrouvons en effet un macro-accuracy average de 0.74.

4 Discussion/Conclusion

La neutralité souvent mise en doute du New York Times semble réelle. Toutefois, on observe des différences entre les sentiments exprimés par les titres des articles et ceux de ses lecteurs, ce qui peut expliquer le biais de perception de la vie politique aux USA.

[4] Lien Collab
[5] Lien Cellule Colab
[6] Lien Cellule Colab
[7] Lien Cellule Colab

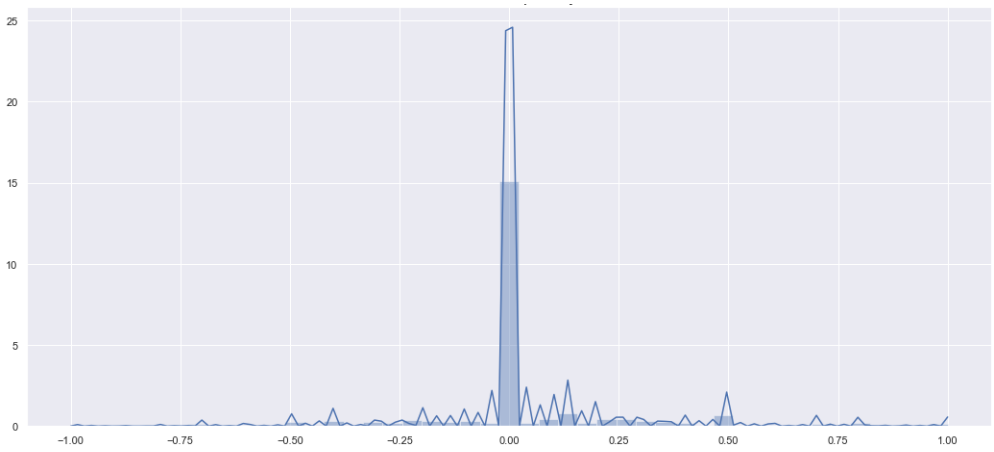

Figure 1: Distribution de la sentimentalité des titres d'articles en 2017

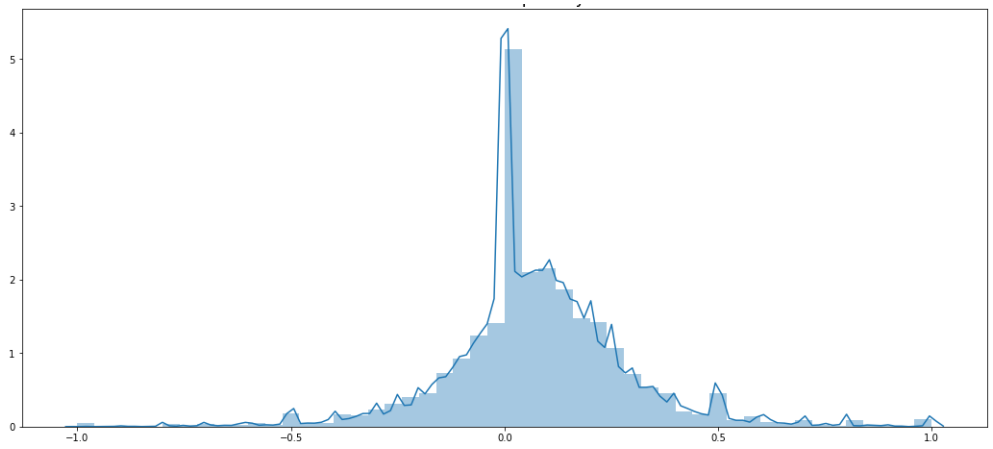

Figure 2: Distribution de la sentimentalité des commentaires d'articles en 2017

Figure 3: Mots les plus courants dans les titres d'articles en 2017

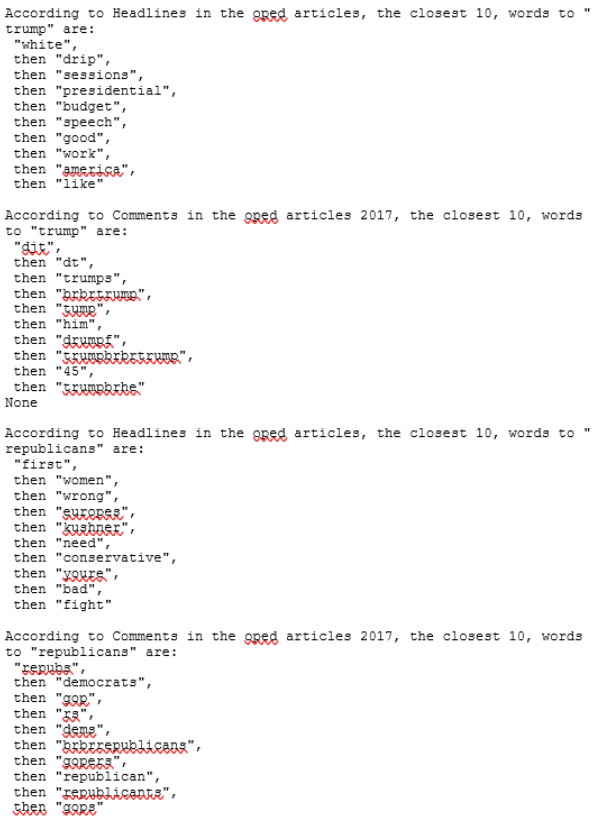

Figure 4: Score de similarité pour les articles éditoriaux de 2017

	word 1	qualitative word	similarity according to headlines model	similarity according to comments model
0	trump	good	0.028188	0.058773
1	trump	bad	0.209827	0.178761
2	republicans	good	0.086315	0.048750
3	republicans	bad	0.030771	0.091754
4	democrats	good	0.126769	0.063767
5	democrats	bad	0.019496	0.092929
6	congress	good	NaN	0.015500
7	congress	bad	NaN	0.026323
8	tax	good	0.009764	0.011650
9	tax	bad	0.132787	0.022053
10	immigration	good	-0.092085	-0.051628
11	immigration	bad	-0.029079	0.026138

Figure 5: Mots et associations: titres et commentaires des articles éditoriaux de 2017

	Cluster #0	Cluster #1	Cluster #2	Cluster #3	Cluster #4	Cluster #5	Cluster #6	Cluster #7	Cluster #8	Cluster #9
1	titletheroguerevolutionistcom	afrin	misconceptions	petulance	egomaniacal	uplifted	herman	afternoons	chemotherapy	unscrupulousness
2	target_blanktheroguerevolutionistcoma	skorea	mindsets	uncouth	vulgarian	relatable	gregory	evenings	nsaids	brbrdjts
3	66a	daesh	methodologies	intemperate	toadying	selfconscious	chaney	430	pulmonary	looktheotherway
4	examplesbra	pakistans	groupings	vitriolic	egodriven	melancholy	schwerner	oclock	hypertension	selfperceived
5	kooh	expansionist	fragmentation	insincere	kleptocrat	blissful	peters	workday	ssri	relentlessness
...	...	...	...	...	...	...	...	...	...	...
96	annbrbrthe	alassads	axioms	horrifyingly	puppeteer	luckiest	barton	yearbrbra	mitochondrial	fideszs
97	livie	afghans	personalized	boastful	godkingemperor	fatalistic	campbell	vacationing	cervical	chameleonlike
98	maude	natos	superficiality	debauched	apparatchik	selfsacrificing	demint	increments	angiogram	surprisetrump
99	fentons	guarantor	purposebrbrthe	childishly	genuflect	indescribable	nichols	monthlong	embolism	lewdness
100	karr	shiites	persuasively	egotist	shyster	despondent	jenny	siena	bowel	powermad

100 rows × 50 columns

Figure 6: Cluster généré à partir de l'analyse Word2Vec des commentaires sur les articles types News de 2018