Actes de la 29e Conférence sur le Traitement Automatique des Langues Naturelles
Avignon, France, 27 juin au 1er juillet 2022

Identification des indicateurs linguistiques de la subjectivité les plus efficaces pour la classification d'articles de presse en français.

Louis Escouflaire
ORM, Ruelle de la Lanterne Magique 14, 1438 Louvain-la-Neuve, Belgique
CENTAL, Place Montesquieu 3, 1438 Louvain-la-Neuve, Belgique
Université catholique de Louvain (UCLouvain)
louis.escouflaire@uclouvain.be

RESUME

Les articles de presse peuvent être répartis en deux genres principaux : les genres de l'information et les genres de l'opinion. La classification automatique d'articles dans ces deux genres est une tâche qui peut être effectuée à partir de traits et mesures linguistiques également utilisées pour l'analyse de la subjectivité. Dans cet article, nous évaluons la pertinence de 30 mesures issues de travaux antérieurs pour la classification d'articles d'information et d'opinion en français. A l'aide de deux modèles de classification différents et à partir d'un échantillon de 13 400 articles publiés sur le site web de la Radio-Télévision Belge Francophone (RTBF), nous avons identifié 18 mesures morphosyntaxiques, lexicosémantiques et stylométriques efficaces pour distinguer les articles plutôt factuels des articles subjectifs.

ABSTRACT

Identifying the most efficient linguistic features of subjectivity for French-speaking press articles classification.

Press articles can be divided into two main genres: information and opinion genres. The automatic classification of articles into these two genres is a task that can be performed using linguistic measures that are also used for subjectivity analysis. In this paper, we evaluate the relevance of 30 measures from previous research for the classification of news and opinion articles in French. Using two different classification models and a sample of 13,400 articles published online by the Belgian Radio-Television of the French Community (RTBF), we identify 18 effective morphosyntactic, lexicosemantic and stylometric measures for distinguishing factual articles from subjective articles.

MOTS-CLES : Analyse de subjectivité, classification automatique de textes, journalism studies, linguistique de corpus.

KEYWORDS: Subjectivity analysis, automatic text classification, journalism studies, corpus linguistics.

Volume 2 : 24e Rencontres Etudiants Chercheurs en Informatique pour le TAL (RECITAL), pages 69–82.

1 Introduction

De nos jours, la diffusion de l'information en ligne passe pour une grande partie des utilisateurs par Facebook, Twitter ou encore Instagram. Les algorithmes de recommandation utilisés par ces médias sociaux enferment les utilisateurs dans des "bulles de filtre", des espaces virtuels au sein desquels ils sont confrontés constamment aux mêmes opinions et où les articles suscitant l'engagement émotionnel du public sont plus souvent mis en avant (Dufrasne & Philippette, 2019). Dans ce contexte, la Radio-Télévision Belge Francophone (RTBF) souligne "l'importance croissante d'avoir des marqueurs identifiables pour différencier les genres de l'information et les genres de l'opinion"[1]. Pour faciliter cette distinction, des outils de traitement automatique du langage peuvent être utilisés.

L'analyse automatique de la subjectivité est un champ du TAL qui a été approché sous différents angles, et qui est parfois considérée comme une sous-tâche de la fouille d'opinion. Suivant les applications, elle consiste à identifier automatiquement dans un corpus donné des textes, des phrases ou des éléments subjectifs, c'est-à-dire dont le contenu est influencé par l'opinion de leur auteur (Wiebe et al., 2004). Pour la détection automatique de la subjectivité dans des textes journalistiques, quelques approches ont déjà été proposées, principalement en anglais (Krüger et al., 2017 ; Alhindi et al., 2020). La question n'a été que peu explorée en français, et encore moins sur des corpus de taille conséquente (Vernier et al., 2009 ; Todirascu, 2019). Dans cet article, nous proposons d'évaluer l'efficacité de 30 indicateurs linguistiques de la subjectivité pour la classification en français d'articles subjectifs et factuels, par le biais d'une analyse d'articles publiés entre 2008 et 2022 par la RTBF. Ce travail s'inscrit dans un projet plus large impliquant l'utilisation de méthodes *deep learning* pour la classification d'articles subjectifs, et l'apport de la classification par traits linguistiques pour en améliorer l'explicabilité.

L'état de l'art présente les discussions liées à la subjectivité, un concept envisagé de différentes manières dans les domaines de la linguistique et du journalisme. Les multiples indicateurs linguistiques de la subjectivité qui ont déjà été évalués dans d'autres études sont également détaillés. Ensuite, la méthodologie appliquée pour l'expérience est décrite dans la section 3, ainsi que le corpus utilisé. Les résultats de la classification sont présentés dans la section 4.

2 État de l'art

2.1 La subjectivité en linguistique

La question de la place occupée par le sujet énonciateur dans le discours, longtemps mise de côté depuis les fondations de Saussure, devient centrale dans la linguistique des années 1970. A la suite des travaux d'Émile Benveniste (1966) sur la Théorie de l'Énonciation, qui refuse la conception classique du langage comme simple outil de communication en déclarant la subjectivité comme

[1] https://www.rtbf.be/article/quand-la-rtbf-donne-son-avis-quelle-place-pour-l-opinion-et-l-edito-dans-nos-medias-10648032?id=10648032, consulté le 16 février 2022.

inhérente au langage, plusieurs auteurs tournent leur attention vers l'instance d'énonciation du discours, et donc vers la subjectivité.

Catherine Kerbrat-Orecchioni (1980) est la première à tenter de faire l'inventaire des traces énonciatives qui peuvent apparaître dans le discours en français. Elle élargit la problématique de l'énonciation de Benveniste en cherchant à recenser les différents types d'unités subjectives, les « procédés linguistiques par lesquels le locuteur imprime sa marque à l'énoncé, s'inscrit dans le message (implicitement ou explicitement) et se situe par rapport à lui » (Kirakossian, 2015). Kerbrat-Orecchioni identifie deux classes principales d'unités subjectives. Les déictiques, d'une part, déjà mis en avant par Benveniste, sont les unités linguistiques dont le sens varie en fonction du sujet, de l'objet et de la situation d'énonciation. Il s'agit principalement de pronoms personnels (ex. *je*, *tu*, *nous*) et d'éléments linguistiques exprimant la localisation temporelle ou spatiale (ex. *hier*, *là-bas*, *ceci*). Les subjectivèmes, d'autre part, sont tous les mots qui portent une connotation évaluative, qui amènent une évaluation ou un jugement affectif de la part du sujet sur l'objet ou sur l'énoncé lui-même. Cette catégorie contient des substantifs (*ânerie*), des adjectifs (*fabuleux*), des verbes (*détester*) et des adverbes (*heureusement*). Cette décomposition de la subjectivité linguistique en unités discursives a été reprise dans de nombreux travaux cherchant à analyser la présence du locuteur dans différentes formes de discours.

Dans le champ du traitement automatique du langage, la fouille d'opinion est une tâche complexe dont l'objectif est d'extraire automatiquement d'un texte le point de vue général qui s'en dégage, le positionnement de l'auteur ou d'une source par rapport à l'objet traité. Une première étape importante dans l'analyse automatique de l'opinion est l'identification dans le texte des faits et des opinions, qui passe d'abord par l'identification des unités subjectives du texte (Eensoo et al., 2011). Cette vision de la subjectivité linguistique comme composite est mise en avant dans les travaux de Janyce Wiebe, qui évaluent la subjectivité d'une phrase à partir de la détection d'*éléments potentiellement subjectifs* (EPS) : « Objective sentences are sentences without significant expressions of subjectivity » (Wiebe et al., 2004). Puis, le taux de subjectivité d'un texte peut être mesuré comme le rapport entre le nombre de phrases subjectives (qui contiennent au moins un "EPS") et le nombre total de phrases dans le texte. L'analyse automatique de la subjectivité à partir d'unités linguistiques peut cependant se réaliser autant au niveau de la phrase qu'au niveau du document (Alhindi et al., 2020). D'autres travaux, présentés dans la section suivante, évaluent le caractère plutôt factuel ou plutôt subjectif d'un texte sur base de différents indicateurs (ou mesures) linguistiques.

2.2 Les mesures linguistiques de la subjectivité

Plusieurs chercheurs se sont attelés à mesurer l'efficacité de nombreux types d'unités subjectives dans différentes langues et dans différents genres de discours. Dans le contexte de cet article sur la subjectivité dans les articles de presse francophone, cette section cherche à dresser une liste (non exhaustive) des traits et mesures linguistiques de la subjectivité les plus efficaces pour les textes du genre journalistique, principalement en langue française.

Concernant les unités subjectives sur le plan morphosyntaxique, la fréquence des déictiques, et en premier lieu des pronoms personnels de la première personne (au singulier et au pluriel), apparaît régulièrement comme un indice pertinent de la présence de l'auteur (Ho-Dac & Küppers, 2011). Le pronom *on* est également considéré comme un bon indicateur de subjectivité dans les articles en français, puisqu'il replace à la fois l'auteur et le lecteur dans le contexte d'énonciation (Todirascu, 2019), tout comme les pronoms personnels de la deuxième personne (Vernier et al., 2009 ; Krüger et al., 2017). Dans leur analyse de plusieurs indicateurs à travers différents médias belges francophones, Ho-Dac & Küppers (2011) attestent de la fréquence élevée des adverbes modalisateurs (ex. *sans doute*, *malheureusement*) dans les articles d'opinion, déjà annoncée par Kerbrat-Orecchioni (1980). Todirascu (2019), en cherchant à classifier des articles d'information et d'opinion à partir de descripteurs linguistiques, observe l'efficacité de plusieurs indices morphosyntaxiques pour identifier les articles d'opinion. On trouve parmi ceux-ci la présence en nombre de conjonctions complexes (ex. *cependant*, *toutefois*), de pronoms relatifs ou encore d'adjectifs. Sur l'anglais, Regmi et Bal (2015) observent également une distribution inégale des parties du discours (*parts-of-speech*) entre les genres journalistiques : les articles d'opinion comptent proportionnellement plus d'adjectifs (Yu & Hatzivassiloglou, 2003), et les articles d'information en moyenne plus de verbes, ce qui peut s'expliquer par l'intuition que les adjectifs apparaissent plus souvent dans des contextes évaluatifs, et que des textes plus factuels se concentrent sur la présentation d'actes avérés. Enfin, une fréquence importante des mots de négation (ex. *ne*, *sans*, *aucun*) semble être un des meilleurs prédicteurs pour la détection automatique d'articles d'opinion en anglais (Krüger et al., 2017 ; Alhindi et al., 2020).

La subjectivité linguistique peut aussi se matérialiser au niveau lexicosémantique, c'est-à-dire au travers des mots utilisés dans le texte et de leurs connotations. D'abord, certains termes de toutes classes grammaticales possèdent (ou peuvent posséder selon le contexte) une valeur axiologique, qu'il s'agisse d'un jugement esthétique (*enlaidir*), moral (*bien*), pragmatique (*important*) ou affectif (*plaisir*) (Vernier et al., 2009). L'utilisation de lexiques d'opinion et de sentiment, des ressources électroniques qui cherchent à regrouper tous les mots subjectifs du vocabulaire d'une langue, est régulièrement montrée comme une manière efficace d'analyser la subjectivité dans le discours (Wilson et al., 2005 ; Todirascu, 2019). Pour le français, on trouve entre autres le lexique FEEL (Abdaoui et al., 2017), adaptation du lexique anglophone NRC EmoLex (Mohammad & Turney, 2013), et le lexique de valence émotionnelle issu de la base de données *Lexique 3* (New et al., 2004 ; Gobin et al., 2017).

Un troisième groupe d'indicateurs linguistiques de la subjectivité concerne la stylométrie. La mesure de la longueur moyenne des mots, des phrases, des citations ou même des articles peut, suivant le corpus étudié, contribuer à la performance de la classification (Todirascu, 2019 ; Alhindi et al., 2020). Différents signes de ponctuation ont également été identifiés comme des indices de subjectivité pertinents. D'abord, les articles d'opinion francophones contiennent plus de points d'interrogation, de points d'exclamation et de points-virgules (Todirascu, 2019). Chaput (2019) a aussi montré l'efficacité de quatre signes de ponctuation particuliers pour l'analyse de la subjectivité dans des textes journalistiques en français du Québec : les points de suspension, les parenthèses, les tirets de mise en relief et les guillemets de mise à distance. Enfin, dans leur analyse de 30 indicateurs

linguistiques de la subjectivité sur un corpus d'articles de presse américains, Krüger et al. (2017) trouvent que le nombre de chiffres utilisés (ex. dates, montants, pourcentages), ainsi que le nombre de virgules, est proportionnellement plus grand dans les articles d'opinion, et que la complexité lexicale est plus élevée dans les articles d'information.

Parmi ces indicateurs linguistiques qui ont fait leur preuve pour l'analyse de la subjectivité ou pour la classification automatique de textes subjectifs, certains n'ont pas encore démontré leur efficacité sur le français ou n'ont pas été évalués sur un large corpus d'articles de presse. Dans cet article, nous cherchons à examiner l'efficacité de 30 mesures pour la classification automatique de textes journalistiques d'opinion et d'information en français.

2.3 La subjectivité en journalisme

Si l'analyse de la subjectivité est depuis longtemps un terreau de recherche fertile en linguistique, cette problématique, ou plutôt celle de l'objectivité, suscite également depuis des décennies des interrogations propres au champ des *journalism studies*. Dès la fin du XIXe siècle aux États-Unis, la présence de l'opinion dans la presse est vue comme un outil commercial exploité par la presse à sensation (Philibert, 2018). On attend du journaliste qu'il poursuive un idéal de vérité à la fois *empirique* et *morale*. Cette dynamique entre la nécessité pour le journaliste de « paraître neutre d'un point de vue politique » et d'être « engagé du point de vue de la morale sociale » reste encore aujourd'hui une préoccupation importante dans les salles de rédaction (Charaudeau, 1997), en particulier pour les médias du service public, comme France Télévisions, la BBC, ou la RTBF en Belgique francophone.

Au cours du temps, l'évolution de la profession journalistique a vu naître différents genres, qui occupent chacun une place plus ou moins importante en fonction des médias. Les genres de l'information (ex. brève, reportage, bulletin météo) et les genres de l'opinion (ex. éditorial, chronique, caricature) se distinguent d'abord par leur place dans le journal ou sur le site internet, par la présence de signaux visuels et de rubriques qui permettent aux lecteurs de les repérer facilement, mais répondent aussi à des normes et des contraintes linguistiques (Grosse, 2001). Pour les genres de l'information, une de ces normes est l'objectivité ; néanmoins, il est communément admis que la recherche de l'objectivité totale est un idéal inaccessible en journalisme, et ce pour plusieurs raisons (Steensen, 2017). Premièrement, le travail essentiel de sélection et de hiérarchisation des informations, tant au sein d'un article qu'au niveau de la salle de rédaction, rend l'impartialité impossible (Gauthier, 1991). Choisir quels faits présenter et dans quel ordre est une tâche par essence subjective mais inévitable en journalisme, l'exhaustivité et la complétude des descriptions étant inatteignables (Philibert, 2018). Ensuite, l'égalité dans la sélection des sources et dans la représentation accordée aux différents points de vue qui peuvent être étayés dans un article est également difficile à maintenir (Koren, 2004). Sur le plan linguistique, la description des faits est indissociable de l'interprétation personnelle de ces faits par le journaliste, guidée par sa vision du monde et par ses expériences. L'impossibilité de l'objectivité est inhérente au fait que la réalité n'est jamais vécue ou perçue de la même manière par deux êtres humains différents. Ainsi, le choix des mots, leur agencement et le style utilisé pour présenter les faits sont autant de facteurs

nécessaires à l'élaboration d'un article qui empêchent la neutralisation de l'influence des opinions de l'auteur sur son texte (Rabatel, 2013). Donc, abandonnant cet idéal de l'objectivité absolue, les journalistes des genres d'information sont contraints d'user de plusieurs procédés linguistiques pour dissimuler leur subjectivité derrière ce que Charaudeau (2006) appelle le « masque de l'effacement énonciatif ». Cette « désubjectivisation » du texte passe par divers effets d'objectivité qui brouillent ou atténuent, sans pour autant la faire disparaître, l'influence de l'opinion du journaliste sur son article (Koren, 2004). Par exemple, l'emploi du pronom indéfini *on*, de tournures impersonnelles, ou encore d'un discours rythmé par les énumérations sont divers moyens de donner à l'article une apparence objective et d'accomplir ce que Tuchman (1972) appelle le « rituel stratégique de l'objectivité ».

Dans le contexte actuel où la diffusion de l'information se fait pour une grande partie de la population à travers les médias sociaux, la question de la place de l'opinion dans les articles de presse revêt une importance accrue. Les médias d'information sont encouragés par le fonctionnement des algorithmes de recommandation à produire du contenu favorisant l'engagement émotionnel, positif ou négatif, des lecteurs (Koivunen et al., 2021). Face à cette hausse constatée de la subjectivité dans le journalisme d'information, il est nécessaire de trouver des moyens d'examiner ce qui se trouve derrière le masque de l'objectivité (Alhindi et al., 2020). Dans la section suivante, nous présentons le corpus et la méthodologie utilisés pour identifier les traits et mesures linguistiques les plus pertinents pour la classification d'articles subjectifs et factuels.

3 Méthodologie

3.1 Corpus

Les échantillons d'articles utilisés pour l'identification des mesures linguistiques de la subjectivité sont issus d'un corpus contenant plus de deux millions d'articles de presse publiés entre 2008 et 2022 par la RTBF, le média de service public belge francophone. Plusieurs centaines d'articles web sont publiés chaque jour sur le site officiel de l'entreprise. Des discussions sont en cours pour rendre le corpus disponible à des fins de reproductibilité des résultats. Les articles du corpus RTBF sont accompagnés de nombreuses métadonnées : titre, auteur, date de publication, catégorie, mots-clés, nombre de vues par jour, etc. Les articles sont répartis dans plusieurs canaux accessibles sur le site web (ex. *info*, *culture*, *sport*, *TV*), puis en catégories plus précises. Parmi les catégories qui peuvent être attribuées aux articles, deux regroupent les articles qu'on peut qualifier d'articles d'opinion : la catégorie *chroniques* (qui reprend des critiques d'œuvres culturelles et des adaptations textuelles des chroniques radio de la RTBF) et la catégorie *opinions* (qui comporte des éditoriaux et des billets d'humeur publiés par des journalistes de la RTBF ainsi que des cartes blanches accordées à des lecteurs ou des spécialistes). Pour éviter de trop grands écarts en termes de variation thématique au sein des données d'entraînement et de test, il a été décidé de se focaliser uniquement sur les articles du canal *info* (rubriques *monde*, *Belgique*, *régions*, *société* et *économie*) pour les articles d'information, et sur les articles d'opinion du canal *info* (rubriques *chroniques* et *opinions*), excluant ainsi par exemple les critiques cinéma et les commentaires sportifs. De cette manière, les articles

d'information et les articles d'opinion conservés pour l'expérience traitent de sujets similaires, ce qui nous permet de nous concentrer sur les aspects discursifs.

Une fois les données filtrées, 6700 articles d'opinion sont conservés. Un ensemble parallèle de 6700 articles d'information correspondant aux critères susmentionnés sont extraits au hasard du corpus entier, formant un total de 13 400 articles. Pour examiner si les deux sous-ensembles contiennent ou non des motifs thématiques trop différents pouvant parasiter la classification, l'algorithme *t-SNE* (Van der Maaten & Hinton, 2008) a été appliqué sur les 13 400 articles. La visualisation du résultat, en Figure 1, montre que les articles d'information recouvrent les thèmes traités dans les articles d'opinion, bien qu'une grande partie des articles d'information comporte des thèmes qui ne sont pas repris dans les thèmes d'opinion (Bogaert et al., 2021). Une étude ultérieure sur un autre corpus permettra de vérifier si nos résultats sont influencés ou non par ce recouvrement partiel des thèmes.

3.2 Mesures

Les indicateurs linguistiques de la subjectivité, qui ont été décrits dans les précédents travaux présentés dans la Section 2.1, sont évalués dans une expérience de classification réalisée sur l'échantillon de 13 400 articles. Ces indicateurs sont répartis en trois types distincts : mesures morphosyntaxiques, lexicosémantiques et stylométriques. Les 31 mesures évaluées par le modèle sont présentées dans le Tableau 1. Les mesures morphosyntaxiques sont calculées à partir de l'article représenté sous la forme d'une liste de tokens ou d'étiquettes morphosyntaxiques (pour *nb_vb* et *nb_adj*), une fois les signes de ponctuation retirés. Les mesures sont normalisées selon la longueur de l'article en calculant la proportion d'éléments significatifs pour la mesure par rapport au nombre total de tokens dans l'article. Par exemple, pour mesurer *nb_neg* sur un article, nous calculons le nombre de mots de négation dans l'article (*non, ni, ne, n, aucun, sans, nul, nulle*) par rapport au nombre total de mots. Concernant les mesures lexicosémantiques, nous calculons le nombre de tokens lemmatisés de l'article qui apparaissent dans le lexique de sentiment correspondant (*Lexique 3*, *FEEL* ou *NRC*), normalisé par le nombre total de mots contenus dans l'article. Les mesures *textblob_subj* et *textblob_sent* sont obtenues en appliquant les fonctions d'analyse de subjectivité et de polarité de la version française de *TextBlob*, un module TAL *Python* multitâche qui se base sur son propre lexique de sentiment (Loria, 2018). Enfin, les mesures stylométriques prenant en compte la ponctuation sont calculées par rapport au nombre total de signes de ponctuation de même niveau dans l'article. Par exemple, *nb_pointvirg* est normalisé en fonction du compte de signes de ponctuation finaux (en fin de phrase) dans le texte. Les mots longs sont les mots de l'article qui comptent au minimum 9 lettres (Todirascu, 2019). Une autre mesure de complexité lexicale utilisée est le *Type-Token Ratio* (TTR) corrigé de Carroll (1964), moins sensible à la longueur du texte que le TTR classique.

3.3 Expérience de classification

Avant l'expérience, chacun des 13 400 articles de l'échantillon est tokenisé, lemmatisé et étiqueté selon la catégorie morphosyntaxique de chaque mot, afin de permettre le calcul des différentes mesures linguistiques. La tâche de classification mise en œuvre dans cet article est fortement

inspirée de la méthodologie de Krüger et al. (2017), qui ont réalisé une recherche similaire sur un corpus d'articles en anglais en analysant l'efficacité de 28 traits linguistiques pour la classification d'articles d'information et d'opinion. Nos 30 indicateurs linguistiques, présentés dans le Tableau 1, sont calculées pour tous les articles, chaque article étant ainsi représenté sous la forme d'un vecteur de 30 caractéristiques. Ensuite, deux modèles de classification différents sont entraînés et utilisés pour identifier les variables les plus efficaces pour la distinction des articles d'information et d'opinion : l'analyse discriminante linéaire (LDA ; Fisher, 1936) et la régression logistique binomiale (LR ; Cox, 1958). La catégorie *information* ou *opinion* attribuée à chaque article par la RTBF elle-même est utilisée comme variable de classification. Une validation croisée stratifiée à 10 blocs est appliquée pour estimer la fiabilité des modèles et pour examiner la potentielle variance induite par les données.

Type de mesure	**Nom de la mesure**	**Description**
Morphosyntaxique	pron_1	Prop. de pronoms et déterminants de la 1e p.
	pron_2	Prop. de pronoms et déterminants de la 2e p.
	pron_1_2	Prop. de pron. et dét. de la 1e/2e p.
	pron_rel	Proportion de pronoms relatifs
	adv_mod	Proportion d'adverbes modalisateurs
	conj_comp	Proportion de conjonctions complexes
	nb_vb	Proportion de verbes
	nb_adj	Proportion d'adjectifs
	nb_neg	Proportion de mots de négation
Lexicosémantique	lexique3_sentiment	Prop. de mots à valence pos/nég. (Lexique3).
	feel_sentiment	Proportion de mots d'émotion (FEEL)
	nrc_sentiment	Proportion de mots d'émotion (NRC)
	textblob_subj	Score moyen de subjectivité selon *TextBlob*
	textblob_sent	Score moyen de sentiment selon *TextBlob*
Stylométrique	length_sentences_chars	Longueur moy. des phrases en nb. caractères
	length_sentences_words	Longueur moy. des phrases en nb. mots
	length_words	Longueur moy. des mots en nb. caractères
	length_citations	Longueur moy. des citations en nb. caractères
	nb_interrog	Proportion de points d'interrogation
	nb_exclam	Proportion de points d'exclamation
	nb_virgules	Proportion de virgules
	nb_pointvirg	Proportion de points-virgules
	nb_deuxpoints	Proportion de deux-points
	nb_susp	Proportion de points de suspension
	nb_tirets	Proportion de tirets (≠ traits d'union)
	nb_parenth	Proportion de paires de parenthèses
	nb_citations	Proportion de texte entre guillemets
	nb_long_words	Proportion de mots longs (8 lettres ou plus)
	nb_digits	Proportion de chiffres
	cttr	Type/Token Ratio corrigé de Carroll

TABLEAU 1. Description des 30 mesures de subjectivité évaluées dans l'expérience.

4 Résultats et discussion

D'abord, l'analyse discriminante linéaire a été appliquée pour la classification des articles, selon la variable *information/opinion* et sur base des 30 mesures décrites dans la section 3.2. Cette première expérience a été effectuée avec plusieurs combinaisons des différents types de traits linguistiques. Pour calculer l'exactitude des prédictions (*predictive accuracy*), que nous appellerons la précision de classification (le nombre d'articles correctement classifiés par le modèle par rapport au nombre total d'articles classifiés), l'échantillon a été réparti en un jeu d'entraînement (90% des articles) et un jeu de test (10%). La précision de classification obtenue sur le jeu de test pour chaque combinaison est présentée dans le Tableau 2.

Analyse discriminante linéaire (LDA)	
Mesures utilisées	**Précision**
Morphosyntaxiques	0.788
Lexicosémantiques	0.685
Stylométriques	0.910
Morphosyntaxiques + lexicosémantiques	0.796
Morphosyntaxiques + stylométriques	0.916
Lexicosémantiques + stylométriques	0.908
Toutes les mesures	0.916

TABLEAU 2. Résultats de la classification des 13 400 articles par LDA.

Utilisées seules, ce sont les mesures stylométriques qui apparaissent comme les plus performantes pour la classification *information-opinion* (précision de 91%), suivies par les mesures morphosyntaxiques (78.8%). Les indicateurs lexicosémantiques obtiennent la précision de classification la plus faible (68.5%). La combinaison de traits qui donne la précision la plus élevée est la combinaison des traits morphosyntaxiques et stylométriques (91.6%). En la comparant avec la classification utilisant les trois ensembles de mesures, on constate que l'ajout des indicateurs lexicosémantiques ne semble pas améliorer la précision de classification.

Pour mettre ces premiers résultats en perspective, une deuxième expérience de classification est réalisée en entraînant un modèle LR sur toutes les mesures calculées à partir des 13 400 articles. L'évaluation de ce modèle est combinée avec une validation croisée stratifiée à 10 blocs, où la répartition en jeux d'entraînement et de test est de 90/10%. La précision de classification moyenne obtenue par le modèle en intégrant les 30 mesures présentées dans le Tableau 1 est de 0.865 (écart-type = 0.014), moins élevée que pour la classification par LDA. Afin d'améliorer la performance de la classification, nous entraînons un nouveau modèle similaire mais dépourvu des 12 indicateurs dont le caractère prédictif n'est pas considéré comme significatif ($p > 0.05$) par le modèle initial. Les 18 mesures conservées et les coefficients de régression qui leur sont attribués par ce nouveau modèle épuré sont présentés dans le Tableau 3. Les mesures à gauche sont plus élevées dans les articles d'opinion, celles à droite dans les articles d'information. L'interprétation du R^2 de McFadden (1973) suggère que les données de l'échantillon sont correctement ajustées à ce modèle ($R^2 = 0.682$). À la suite de cette étape de sélection des caractéristiques, nous obtenons une précision

de classification de 0.897 (écart-type = 0.008) avec le nouveau modèle LR à 18 mesures, plus efficace que le modèle LR utilisant les 30 mesures.

Opinion				Information			
Mesure	**coef.**	**SE**	***p***	**Mesure**	**coef.**	**SE**	***p***
nb_on	101.3	7.32	< 0.001	*nb_digits*	79.26	3.36	< 0.001
pron_rel	54.34	4.49	< 0.001	*pron_1*	46.28	4.95	< 0.001
nb_neg	49.51	5.31	< 0.001	*nb_vb*	13.2	1.25	< 0.001
nb_adj	22.85	1.62	< 0.001	*length_words*	3.72	0.08	< 0.001
lexique3_sentiment	16.34	5.78	0.005	*nb_citations*	0.37	0.01	< 0.001
nrc_sentiment	10.39	1.71	< 0.001				
nb_exclam	9.35	0.76	< 0.001				
nb_interrog	6.7	0.63	< 0.001				
nb_pointvirg	6.22	0.9	< 0.001				
nb_susp	3.90	0.79	< 0.001				
nb_deuxpoints	3.74	0.48	< 0.001				
blob_sent	2.43	0.59	< 0.001				
cttr	1.85	0.04	< 0.001				

TABLEAU 3. Modèle de classification par régression logistique (LR) à 18 prédicteurs ($R^2 = 0.682$).

Les courbes d'apprentissage des deux modèles LR, calculées en entrainant le modèle sur 30 paliers successifs de taille d'échantillon (de 0.5% à 100% des données d'entrainement) et en le mesurant sur un même jeu de test (20% des données) avec validation croisée (k = 10), sont présentées dans la Figure 2. On peut observer que la précision de classification des deux modèles croit significativement jusqu'à atteindre une taille d'échantillon d'entraînement de 1000 articles, avant de se stabiliser. On peut néanmoins constater que la précision du modèle basé sur les 18 mesures continue à croître légèrement mais de manière constante avec la taille du jeu d'entraînement, tandis que la précision du modèle utilisant les 30 mesures est moins stable et moins élevée. Cette instabilité est probablement due au bruit provoqué dans la classification par les 12 mesures considérées comme non prédictives par le modèle. Cette comparaison des courbes d'apprentissage nous permet de confirmer que certaines mesures parmi les 30 tirent la précision de classification vers le bas, et qu'il est plus efficace de ne conserver que les mesures les plus significatives d'après le modèle.

Dans le Tableau 3, on observe que les mesures avec le plus haut coefficient de régression sont principalement des mesures morphosyntaxiques et stylométriques (*nb_on*, *nb_digits*, *pron_rel*), ce qui confirme les résultats de la classification par LDA présentés dans le Tableau 2. Sur les 18 mesures conservées dans le modèle utilisant la LR, 9 sont des mesures stylométriques, s'alignant avec la performance de ces mesures pour la classification par LDA. De même, seules trois mesures lexicosémantiques sont significatives pour le modèle LR, avec un coefficient de régression en moyenne peu élevé. Nous utilisons les résultats obtenus par le modèle à 18 mesures pour examiner de manière plus détaillée quelles sont les mesures les plus discriminantes pour la classification d'articles d'information et d'opinion.

Les mesures morphosyntaxiques auxquelles le modèle LR accorde le plus de poids sont d'abord les pronoms et déterminants. La fréquence du pronom *on*, en premier, est le meilleur prédicteur des articles d'opinion (*nb_on*). Il n'apparaît que très rarement dans les articles d'information, ce qui confirme les résultats de Todirascu (2019). La même observation peut être faite pour les pronoms relatifs (*pron_rel*), significativement plus nombreux dans les articles d'opinion. L'ubiquité dans les articles d'opinion du pronom *on*, décrit par Koren (2004) comme l'instrument principal de l'effacement énonciatif du journaliste, peut expliquer le fait que le nombre de pronoms et déterminants de la 1e personne soit considéré par le modèle comme un prédicteur en faveur des articles d'information (*pron_1*), ce qui va cependant à l'encontre des résultats obtenus par Krüger et al. (2017) pour les articles anglophones. Ce résultat pourrait aussi être expliqué par la présence plus fréquente de ces marques de la 1e personne à l'intérieur de citations (la proportion de texte entre guillemets étant plus élevée dans les articles d'information), bien que ces hypothèses doivent encore être confirmées. Ensuite, la fréquence des mots de négation, confirmant cette fois les observations obtenues pour l'anglais par Krüger et al., 2017, est plus élevée dans les articles d'opinion. En ce qui concerne les étiquettes morphosyntaxiques, nos résultats sont en accord avec les études réalisées sur le sujet (Yu & Hatzivassiloglou, 2003 ; Regmi & Bal, 2015) : les articles d'opinion comptent en moyenne plus d'adjectifs (*nb_adj*) et les articles d'information plus de verbes (*nb_vb*).

FIGURE 1 : Visualisation des thèmes des 13 400 articles par t-SNE.

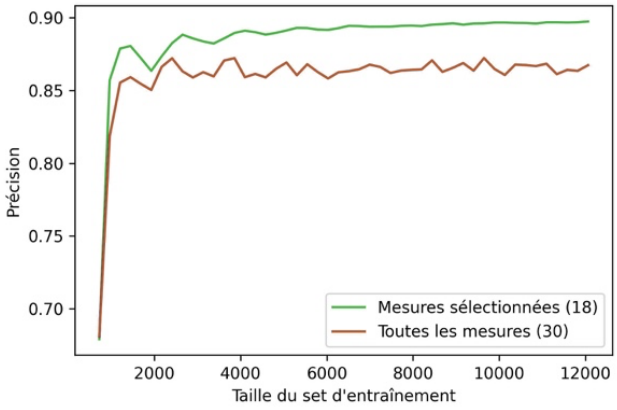

FIGURE 2. Courbes d'apprentissage pour les modèles de classification LR à 30 et à 18 mesures.

Parmi les lexiques d'émotions utilisés pour calculer les mesures lexicosémantiques, le lexique de valence basé sur Lexique 3 est celui auquel le modèle donne le plus de poids, suivi par la traduction française du lexique NRC. Ces résultats sont intéressants, puisque le lexique de valence de Gobin et al. (2017) basé sur Lexique 3 ne comporte que 1.286 entrées, tandis que celui de NRC regroupe 14 000 mots. On peut en conclure que la mesure basée sur les mots portant un sens évaluatif la plus performante pour cette tâche de classification est celle qui ne prend en compte que les mots évaluatifs les plus évidents, et donc un lexique plus restreint. Puis, pour *TextBlob*, seul le score de polarité est considéré comme prédictif, quoique très légèrement, pour la classification.

Concernant les mesures stylométriques, on observe avant tout que la fréquence normalisée des nombres (*nb_digits*) est bien plus élevée dans les articles d'information, ce qui va à l'encontre des résultats de Krüger et al. (2017) pour l'anglais, mais confirme cependant leur hypothèse préliminaire selon laquelle les articles qui mentionnent plus de dates, montants ou pourcentages sont aussi les plus factuels. Ensuite, nos résultats appuient que les fréquences des points d'exclamation (*nb_exclam*), d'interrogation (*nb_interrog*) et des points de suspension, déjà observées par Todirascu (2019), auxquelles nous ajoutons celles des points-virgules (*nb_pointvirg*) et des deux-points (*nb_deuxpoints*), sont des prédicteurs efficaces pour les articles d'opinion. Cependant, la distribution des autres signes de ponctuation proposés par Chaput (2019) comme des indices de subjectivité pertinents, à savoir le nombre de parenthèses et de tirets, ne sont pas considérés par notre modèle comme des mesures significatives pour la classification des articles d'opinion. Enfin, le modèle attribue un faible coefficient au CTTR, plus élevé dans les articles d'opinion. Au contraire, la longueur moyenne des mots (*length_words*) apparaît comme un indicateur de complexité plutôt en faveur des articles d'information.

5 Conclusion

Les résultats de cette expérience ont montré l'efficacité d'un ensemble d'indicateurs linguistiques pour la classification des articles d'information et d'opinion d'un média belge francophone, la RTBF. Les types de mesures qui se sont révélés les plus prédictifs pour cette tâche sont les mesures morphosyntaxiques et stylométriques, les mesures lexicosémantiques étant considérées comme moins importantes par les deux modèles de classification utilisés pour l'expérience. Nos résultats confirment en partie ceux obtenus dans le cadre de recherches antérieures sur le français ou sur l'anglais, bien que certaines constatations divergent de ce qui avait été observé jusque-là dans le domaine.

Dans le cadre d'une étude à venir, les 18 indicateurs identifiés dans cet article seront utilisés pour une analyse diachronique et à plus grande échelle du corpus RTBF, afin d'observer l'évolution de la présence de ces traits linguistiques de subjectivité dans les articles du média publiés entre 2008 et 2022. De plus, des discussions sont en cours avec d'autres groupes de presse pour enrichir le corpus et étendre les analyses à plusieurs médias francophones. De futures perspectives de recherche impliquent également l'évaluation d'autres mesures linguistiques de la subjectivité attestées dans la littérature et dont l'efficacité n'a pas été testée dans cette étude, comme les n-grammes ou l'expression de la modalité. Enfin, il est prévu d'explorer l'importance de prendre en compte (ou non) le contenu de l'article à l'intérieur des citations pour évaluer la subjectivité du texte.

Remerciements

L'auteur tient à remercier particulièrement Jérémie Bogaert, Antonin Descampe, Cédrick Fairon et François-Xavier Standaert, ainsi qu'Alain Guillet du service SMCS de l'UCLouvain, pour leur aide et leur soutien au cours de la réalisation de ce travail.

Références

ABDAOUI A., AZE J., BRINGAY S., & PONCELET P. (2017). FEEL: A French Expanded Emotion Lexicon. *Language Resources and Evaluation*, 51(3), 833-855.

ALHINDI T., MURESAN S., & PREOTIUC-PIETRO D. (2020). Fact vs. Opinion: The Role of Argumentation Features in News Classification. *Proceedings of the 28th International Conference on Computational Linguistics*, 6139-6149. DOI : 10.18653/v1/2020.coling-main.540.

BENVENISTE E. (1966). De la subjectivité dans le langage. *Problèmes de linguistique générale*, Paris, Gallimard (coll. Bibliothèque des sciences humaines), 258-266.

BOGAERT J., CARBONNELLE Q., DESCAMPE A., & STANDAERT F.-X. (2021). Can Fake News Detection be Accountable? The Adversarial Examples Challenge. *41st WIC Symposium on Information Theory in the Benelux*.

CARROLL J. B. (1964). *Language and thought*. Englewood Cliffs, NJ: Prentice-Hall.

CHARAUDEAU P. (1997). *Le discours d'information médiatique : la construction du miroir social*. Nathan.

CHARAUDEAU P. (2006). Discours journalistique et positionnements énonciatifs. Frontières et dérives. *Semen, 22*. DOI : 10.4000/semen.2793.

CHAPUT L. (2019). Sur quelques marques de subjectivité dans le journalisme d'information politique de 1945 à 2015 au Québec. *Mots, 119*, 151-168. DOI : 10.4000/mots.24586.

CHENLO J. M., & LOSADA D. E. (2014). An empirical study of sentence features for subjectivity and polarity classification. *Information Sciences, 280*, 275-288. DOI : 10.1016/j.ins.2014.05.009.

COX D. R. (1958). The regression analysis of binary sequences. *Journal of the Royal Statistical Society: Series B (Methodological), 20(2)*, 215–232.

DUFRASNE M., & PHILIPPETTE T. (2019). Les effets des bulles de filtres ou bulles informationnelles sur la formation des opinions. *Journée d'étude pour le lancement du projet Alg-opinion*.

EENSOO E., BOURION E., SLODZIAN M., & VALETTE M. (2011). De la fouille de données à la fabrique de l'opinion. L*es Cahiers du numérique, Vol. 7(2)*, 15-40.

FISHER R. A. (1936). The Use of Multiple Measurements in Taxonomic Problems. *Annals of Eugenics, 7(2)*, 179-188.

GAUTHIER G. (1991). La mise en cause de l'objectivité journalistique. *Communication, 12(2)*, 80-115.

GOBIN P., CAMBLATS A. M., FAUROUS W., & MATHEY S. (2017). Une base de l'émotionalité (valence, arousal, catégories) de 1286 mots français selon l'âge (EMA). *European Review of Applied Psychology, 67(1)*, 25-42.

GROSSE E.-U. (2001). Evolution et typologie des genres journalistiques : Essai d'une vue d'ensemble. *Semen, 13*. DOI : 10.4000/semen.2615.

HO-DAC L.-M., & KÜPPERS A. (2011). La subjectivité à travers les médias : Étude comparée des médias participatifs et de la presse traditionnelle. *Corpus, 10*, 179-199. DOI : 10.4000/corpus.2076.

KERBRAT-ORECCHIONI C. (2009). *L'énonciation : de la subjectivité dans le langage*. Armand Colin.

KIRAKOSSIAN A. (2015). *La subjectivité linguistique dans l'acceptation de Catherine Kerbrat-Orecchioni*. Université d'État d'Erevan.

KOIVUNEN A., KANNER A., JANICKI M., HARJU A., HOKKANEN J., & MÄKELÄ E. (2021). Emotive, evaluative, epistemic: A linguistic analysis of affectivity in news journalism. *Journalism, 22(5)*, 1190-1206. DOI : 10.1177/1464884920985724.

KOREN R. (2004). Argumentation, enjeux et pratique de l'« engagement neutre » : Le cas de l'écriture de presse. *Semen, 17*. DOI : https://doi.org/10.4000/semen.2308.

KRÜGER K. R., LUKOWIAK A., SONNTAG J., WARZECHA S., & STEDE M. (2017). Classifying news versus opinions in newspapers: Linguistic features for domain independence. N*atural Language Engineering, 23(5)*, 687-707. DOI : 10.1017/S1351324917000043.

LORIA S. (2018). *TextBlob Documentation. Release 0.15, 2*.

MCFADDEN D. (1973). Conditional logit analysis of qualitative choice behaviour. *Frontiers in Econometrics*, 105–142.

MOHAMMAD S. & TURNEY P. (2013). Crowdsourcing a Word-Emotion Association Lexicon, *Computational Intelligence, 29 (3)*, 436-465.

NEW B., PALLIER C., BRYSBAERT M., FERRAND L. (2004) Lexique 2: A New French Lexical Database. *Behavior Research Methods, Instruments, & Computers, 36(3)*, 516-524.

PHILIBERT J.-R. (2018). Fake news, désinformation et autres critiques actuelles du journalisme : Mise en perspective à l'aune de critiques adressées à la presse écrite nord-américaine entre 1870 et 1910. *Journal of the Seminar of Discursive Logic, Argumentation Theory and Rhetoric, 16(2)*.

RABATEL A. (2013). L'engagement du chercheur, entre éthique d'objectivité et éthique de subjectivité . *Argumentation et analyse du discours, 11*. DOI : 10.4000/aad.1526.

REGMI S. & BAL K. B. (2015). What Makes Facts Stand Out from Opinions? *Creativity in Intelligent Technologies and Data Science, Vol. 535*, 655-66. DOI : 10.1007/978-3-319-23766-4_51.

STEENSEN S. (2017). Subjectivity as a Journalistic Ideal. *Putting a Face on it: Individual Exposure and Subjectivity in Journalism*. Cappelen Damm Akademisk, 25-47.

TODIRASCU A. (2019). Genre et classification automatique en TAL : Le cas de genres journalistiques. *Linx, 78*. DOI : 10.4000/linx.3183.

TUCHMAN G. (1972). Objectivity as strategic ritual: An examination of newsmen's notions of objectivity. *American Journal of sociology, 77(4)*, 660-679.

VAN DER MAATEN L., & HINTON G. (2008). Visualizing data using t-SNE. *Journal of machine learning research, 9(11)*.

WIEBE J., WILSON T., BRUCE R., BELL M., & MARTIN M. (2004). Learning Subjective Language. *Computational Linguistics, 30(3)*, 277-308. DOI : 10.1162/0891201041850885.

WILSON T., HOFFMANN P., SOMASUNDARAN S., KESSLER J., WIEBE J., CHOI Y., CARDIE C., RILOFF E., & PATWARDHAN S. (2005). OpinionFinder: A system for subjectivity analysis. *Proceedings of HLT/EMNLP on Interactive Demonstrations*, 34-35.

YU H., & HATZIVASSILOGLOU V. (2003). Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. *Proceedings of the 2003 conference on Empirical methods in natural language processing*, 129-136. DOI : 10.7916/D88W3NN0.