La subjectivité dans le journalisme québécois et belge : Transfert de connaissances inter-médias et inter-cultures

Louis Escouflaire[1], Antonin Descampe[2], Antoine Venant[3], Cédrick Fairon[4]

[1]UCLouvain – louis.escouflaire@uclouvain.be; antonin.descampe@uclouvain.be; antoine.venant@umontreal.ca; cedrick.fairon@uclouvain.be

Abstract

We compare the ability to transfer knowledge across journalistic sources and cultures for an opinion vs. news articles classification task in French, between two models: the large language model CamemBERT and a statistical model based on linguistic features. We use a corpus of 80,000 articles published by eight Canadian and Belgian French media, which also allows us to compare how opinion and news articles contrast in the two media landscapes. Our results show that CamemBERT has a higher knowledge transfer capacity than the feature-based model, and attest to the existence of a difference between Quebec and Belgian press discourse.

Keywords: journalism, transfer learning, Large Language Models, cross-cultural comparison.

Résumé

Nous comparons la capacité de transfert de connaissances dans le cadre d'une tâche de classification d'articles de presse d'opinion et d'information en français, d'un média à l'autre et d'une culture journalistique à l'autre, pour deux modèles : le grand modèle de langage francophone CamemBERT et un modèle statistique basé sur des traits linguistiques. Nous utilisons un corpus de 80 000 articles publiés par huit médias québécois et belges, ce qui nous permet également de comparer la manière dont les articles d'opinion et d'information s'opposent dans les deux paysages journalistiques francophones. Nos résultats montrent que CamemBERT possède une capacité de transfert plus élevée que le modèle statistique, et attestent d'une différence entre les discours de presse québécois et belge.

Mots clés : journalisme, transfert de connaissance, grands modèles de langage, comparaison inter-culturelle.

1. Introduction

La tradition journalistique occidentale a toujours séparé les articles de presse en deux grandes catégories : les articles d'information, qui présentent l'actualité de la manière la plus neutre (ou "objective") possible, et les articles d'opinion, dans lesquels les journalistes partagent ouvertement leur avis "subjectif" sur le sujet dont ils traitent (Grosse, 2001). Cependant, les nouveaux modes de réception de l'information (sur le web et les réseaux sociaux) rendent cette distinction de plus en plus floue pour les lecteurs. Dans le domaine du Traitement Automatique du Langage (TAL), la classification d'articles d'opinion et d'information a longtemps été considérée comme une tâche particulièrement complexe (Ravi et Ravi, 2015). Toutefois, l'avènement des grands modèles de langage a repoussé les frontières de l'état de l'art pour différentes tâches de classification de textes (Devlin et al., 2018). Surpassant les méthodes de classification plus traditionnelles, ces modèles basés sur des réseaux de neurones acquièrent des compétences linguistiques générales en étant pré-entraînés sur de vastes corpus de données brutes. Ils se montrent également capables de transférer leurs connaissances à des tâches connexes à celles sur lesquelles ils ont été peaufinés, par exemple en les généralisant à d'autres langues ou d'autres domaines (Qasim et al., 2022). Cependant, à notre connaissance cette capacité n'a pas encore été évaluée dans la classification d'articles d'opinion et d'information.

Dans cet article, nous évaluons pour cette tâche les capacités de transfert de deux types de modèles de classification : un grand modèle de langage neuronal et un modèle statistique basé sur des traits linguistiques. Dans ce but, nous entraînons nos modèles sur des textes issus d'un média, puis les évaluons sur des textes issus de médias différents. Notre étude s'appuie sur un corpus de 80 000 articles extraits de huit médias québécois et belges francophones, ce qui nous permet d'explorer la capacité des modèles à transférer leurs connaissances non seulement d'un média à l'autre, mais aussi entre deux cultures journalistiques. De plus, une analyse de traits auxquels les modèles statistiques accordent le plus d'importance selon le média sur lequel ils ont été entraînés nous permet d'examiner la manière dont la distinction stylistique entre presse d'opinion et d'information se réalise dans les médias québécois et belges francophones.

2. Etat de l'art

2.1. La classification automatique d'articles d'opinion et d'information

En TAL, la classification automatique de textes subjectifs est une sous-tâche du domaine de l'analyse de sentiments, qui consiste à déterminer si un texte présente des informations objectives ou subjectives, c'est-à-dire s'il relate des faits ou bien s'il contient les opinions, émotions ou sentiments de son auteur ou d'une autre source rapportée par l'auteur (Tang et al., 2009). Appliquée à l'analyse de discours de presse, cette tâche peut être rapprochée de celle consistant à déterminer si un article appartient au genre journalistique de l'opinion (éditoriaux, billets d'humeur, commentaires ou critiques), ou au genre de l'information (dépêches d'agences de presse, nouvelles, actualités). Comme pour d'autres tâches de classification en TAL, les grands modèles de langage basés sur des *transformers* montrent des résultats significativement meilleurs que les méthodes plus traditionnelles utilisant des traits linguistiques (Bogaert et al., 2024). Ces deux approches sont présentées plus en détails dans la section 3.2 de cet article.

Concernant l'utilisation de modèles *transformers*, il est important de souligner que malgré leurs performances élevées, ils souffrent d'une série de limitations. D'abord, le manque d'explicabilité de ces modèles complexes, dits "boîtes noires", reste une préoccupation majeure dans le domaine du TAL, particulièrement dans le cadre de tâches et d'applications dans lesquelles les décisions des modèles et leurs biais peuvent avoir des implications importantes (Lyu et al., 2022). Ensuite, leur coût computationnel et environnemental est très élevé et souvent négligé. Au contraire, les modèles d'apprentissage automatique traditionnels, qui n'utilisent pas de réseaux de neurones profonds (*deep learning*), sont moins complexes et ont la capacité de fournir des explications interprétables et fidèles à leurs prédictions.

2.2. L'objectivité journalistique, de l'Amérique à l'Europe

En journalisme, la notion d'objectivité est perçue comme une norme professionnelle qui se matérialise sous la forme de règles éditoriales comme l'usage des guillemets de citation, la préférence pour un discours neutre, ou l'équilibre des points de vue (Schudson, 2001). Pour éviter tant que possible de produire des articles d'information influencés par leurs choix et leur subjectivité, les journalistes appliquent une variété de mécanismes d'objectivation du discours. Dans le journalisme d'opinion, ouvertement subjectif, le discours des éditorialistes et chroniqueurs est plutôt marqué par des indicateurs de subjectivité, ce qui le distingue du journalisme d'information (Charaudeau, 2006).

Même s'il s'agit de la norme dominante dans le journalisme occidental, l'objectivité s'est manifestée différemment dans les pays anglo-saxons et en Europe continentale (Williams, 2006). Il est communément admis qu'elle est apparue au 19^{e} siècle aux Etats-Unis, à la fois

entraînée par des révolutions technologiques et économiques qui permettaient aux journaux de toucher un plus large public plus rapidement, et aussi à la suite des bouleversements politiques de l'époque et des mouvements idéologiques et culturels liés à l'essor de la doctrine positiviste (Muñoz-Torres, 2012). L'objectivité ne touche l'Europe que bien plus tard, au tournant des années 1920, où elle se heurte à un journalisme européen qui lui est moins perméable : la presse d'opinion occupe une place importante, la publicité n'est pas la bienvenue, et l'influence des milieux littéraires empêche les valeurs américaines de rigueur et de rapidité de s'imposer. Petit à petit, un journalisme tourné vers l'information finit par y émerger au cours du 20e siècle, laissant néanmoins toujours une place à l'élaboration stylistique et à la présence de la voix de l'auteur dans l'article (Chalaby, 1996).

Aujourd'hui, ces différences historiques ont évolué mais continuent à se ressentir. Le modèle américain, caractérisé par la recherche constante d'objectivité et l'indépendance du journaliste, domine le paysage médiatique occidental (Williams, 2006). Malgré l'influence américaine, le modèle journalistique européen maintient encore certaines de ses spécificités, même s'il n'est pas monolithique : Esser et Umbricht (2013) soulignent que le journalisme « méditerranéen » (dans lequel s'inscrit en partie le journalisme français) est plus orienté vers l'opinion et les registres littéraires que le modèle « corporatiste » germanique et d'Europe du Nord, dont les idéaux se rapprochent de plus en plus du modèle américain. Ces variations interculturelles peuvent être cernées plus finement par le biais d'analyses comparatives de contenus journalistiques issus de différentes sociétés (Hallin et Mancini, 2012).

Parmi les cultures dont le modèle journalistique n'a pas encore été étudié dans une visée contrastive, la Belgique francophone et le Québec constituent des cas particulièrement intéressants, qui apparaissent tous deux comme des paysages à cheval entre le modèle anglo-saxon et le modèle méditerranéen. Le paysage médiatique belge est à l'image de la segmentation politique du pays : partagé entre les territoires francophone et néerlandophone, et marqué par les spécificités culturelles qui distinguent chaque communauté linguistique. La tradition flamande emprunte beaucoup aux Pays-Bas, tandis que la proximité culturelle entre la Wallonie et la France n'est plus à démontrer. La Belgique est une monarchie constitutionnelle fédérale, ce qui la rapproche politiquement du modèle corporatiste décrit par Esser et Umbricht (2013). D'autre part, le journalisme québécois est aussi nourri des différents modèles. L'histoire du Québec est marquée par la tension entre les cultures américaine et européenne. Sur le plan géographique, les Québécois sont toujours en contact direct avec les Etats-Unis et le reste du Canada ; sur le plan linguistique, ils restent très attachés à l'héritage de leurs ancêtres français.

3. Méthodologie

3.1. Données

3.1.1. Constitution du corpus

Un total de 80 000 articles de presse ont été rassemblés dans le cadre de ce projet de recherche, au moyen d'outils de moissonnage web (*scraping*) et de requêtes auprès de la base de données Europresse[1]. Le corpus contient des articles publiés par huit médias différents, à raison de 10 000 articles par média. Quatre médias sont québécois, quatre sont belges francophones. Pour chaque média, nous avons rassemblé 5000 articles correspondant aux genres journalistiques de l'opinion et 5000 articles appartenant aux genres de l'information. Ces étiquettes ont été

[1] https://www.europresse.com/

attribuées aux articles par les médias qui les ont publiés. Le corpus complet contient donc une distribution équilibrée de 40 000 articles d'opinion et 40 000 articles d'information.

Pour constituer les sous-corpus, nous avons utilisé la même approche pour chaque média. Nous avons d'abord récolté et tokenizé 5000 articles d'opinion, puis appliqué une allocation de Dirichlet latente (LDA) pour identifier automatiquement les différents thèmes abordés dans chacun de ces articles. La granularité de l'analyse dépend d'un paramètre spécifiant le nombre de thèmes utilisés pour expliquer statistiquement les données. Nous avons ajusté cette valeur de manière à minimiser la perplexité du modèle LDA, en limitant le nombre maximum de thèmes identifiés à dix[2]. Ensuite, nous avons constitué un corpus d'articles d'information en moissonnant des articles issus des rubriques correspondant aux différents thèmes identifiés par la LDA, au prorata de la représentation de chaque thème dans les articles d'opinion. Par exemple, en visualisant les thèmes modélisés par LDA pour les articles d'opinion du média La Presse, nous observons que parmi les 5000 textes, environ 20% traitent d'économie[3]. Pour constituer le corpus d'information, nous extrayons donc 1000 articles de la rubrique Économie de La Presse (et ainsi de suite pour les autres thèmes identifiés). Cette approche nous permet de constituer deux sous-corpus contenant respectivement 5000 articles d'opinion et 5000 articles d'information, de sorte que les thèmes abordés sont distribués de manière aussi équivalente que possible dans les deux sous-corpus. En vue d'entraîner et évaluer les modèles de classification, chaque sous-corpus est réparti en trois ensembles : d'entraînement (8000 articles), de développement (1000) et de test (1000), tous équilibrés entre les classes *opinion* et *information*.

3.1.2. Description

Les huit médias représentés dans le corpus peuvent être caractérisés selon trois dimensions : origine (québécois ou belge), portée (média régional ou national francophone) et visée (généraliste ou sensationnaliste). Les quatre médias québécois représentés dans le corpus sont *La Presse* (national et généraliste), *Le Devoir* (national et généraliste), *Le Journal de Montréal* (*JDM* ; national et sensationnaliste) et *Le Soleil* (régional et généraliste). Les quatre médias belges sont la *RTBF* (national et généraliste), *Le Soir* (national et généraliste), *La Libre Belgique* (national et généraliste) et *L'Avenir* (régional et sensationnaliste). De plus, la *RTBF* est un média de service public, qui répond dès lors à des impératifs éditoriaux potentiellement différents des autres médias dans le corpus. Tous les médias possèdent un site web alimenté en continu et sont distribués quotidiennement au format papier, sauf *La Presse* (depuis 2017), *Le Soleil* (depuis fin 2023) et la *RTBF* qui ne sont accessibles que via le web. Tous les articles dans le corpus ont été publiés entre 2014 et 2023. Le corpus total[4] contient 55.003.306 tokens. Les articles *opinion* contiennent 812 tokens en moyenne, contre 624 pour les articles *information*.

3.2. Modèles de classification

3.2.1. Modèles CamemBERT

Nous utilisons le modèle CamemBERT (Martin et al., 2020) dans sa version de base et insensible à la casse, qui contient 110 millions de paramètres. CamemBERT est fondé sur l'architecture du modèle RoBERTa (Liu et al., 2019) et a été pré-entraîné sur les textes en

[2] Une perplexité basse signifie que le modèle est capable de bien prédire le contenu d'un document à partir de sa structure thématique, et constitue donc un bon indicateur de la qualité de cette dernière.

[3] LDA est une méthode non supervisée qui ne fournit pas d'étiquette pour un thème donné mais une distribution sur les mots. L'attribution des catégories (politique, sport, etc.) est faite manuellement à partir de cette dernière.

[4] A des fins de reproductibilité, le corpus sera mis à disposition sur demande.

français du corpus OSCAR (138 Go de texte). Comme RoBERTa, CamemBERT peut être peaufiné (*fine-tuned*) sur des données additionnelles en vue d'être utilisé pour une tâche spécifique, comme la classification de textes.

Dans cet article, nous peaufinons dix versions différentes de CamemBERT : 8 versions 'mono-médias', chacune peaufinée sur l'ensemble d'entraînement du sous-corpus correspondant à l'un des huit médias, et deux versions 'multimédias', l'une peaufinée sur 8000 articles issus des quatre médias québécois (2000 articles par média), et l'autre peaufinée sur 8000 articles issus des quatre médias belges. Nous appelons ces deux modèles multimédias les modèles *Québec* et *Belgique*. Pour chaque version de CamemBERT, le peaufinage est effectué avec les mêmes hyperparamètres (taille de lot = 4 ; vitesse d'apprentissage = 2×10^{-5}) durant deux époques.

3.2.2. Modèles basés sur des traits

L'autre type de modèle que nous utilisons se base sur un classifieur statistique utilisant trente-trois traits textuels issus de l'état de l'art sur la subjectivité linguistique. La plupart de ces traits reposent sur la présence ou la fréquence de certains tokens ou types de tokens dans le texte à classer. Dix-neuf de ces traits ont déjà été identifiés comme des prédicteurs efficaces de l'opinion dans les articles belges de la *RTBF* (Escouflaire, 2022) : les adjectifs, verbes, pronoms et déterminants des première et deuxième personnes, pronoms relatifs, le pronom indéfini *on*, les signes de ponctuation (points-virgules, deux-points, points d'exclamation, points d'interrogation et points de suspension), guillemets, chiffres, mots de négation, mots apparaissant dans des lexiques de sentiment et de polarité, le nombre de citations dans l'article, la longueur moyenne des tokens et le ratio type-token corrigé. Quatorze traits additionnels sont utilisés pour améliorer les performances du modèle : les adverbes modaux, conjonctions complexes, virgules, hapax, marqueurs de discours, marqueurs temporels absolus et relatifs, verbes à la voix passive, verbes de citation et de pensée, les mots de plus de sept caractères, la concrétude (Bonin et al., 2018), la fréquence subjective et l'imageabilité (Desrochers et Thompson, 2009) moyennes des mots de l'article. Le calcul de la plupart de ces traits se base sur le ratio du nombre de tokens correspondant aux traits par rapport au nombre de tokens total de l'article. Le modèle de classification est basé sur une régression logistique, de sorte que les poids attribués à chaque trait sont ajustés au cours de l'entraînement. Dix versions de ce modèle basé sur des traits sont entraînées sur les différents ensembles d'entraînement (dont des modèles *Québec* et *Belgique*). Pour chaque version du modèle, nous appliquons une recherche par grille pour sélectionner la combinaison d'hyperparamètres optimale pour la régression.

4. Résultats et discussion

4.1. Classification

4.1.1. Modèles CamemBERT

La Figure 1 présente l'exactitude (le nombre de prédictions correctes par rapport au nombre total de prédictions) de chaque modèle CamemBERT sur les dix ensembles de test. La première observation est que, naturellement, les modèles montrent systématiquement une exactitude plus élevée sur l'ensemble de test issu du même média que l'ensemble d'entraînement sur lequel ils ont été peaufinés. L'exception à cette règle concerne l'ensemble de test du *JDM*, sur lequel la plupart des modèles montrent une meilleure exactitude que sur leur ensemble de test correspondant. Une hypothèse d'explication peut être que la différence de style journalistique entre les articles d'opinion et d'information est plus claire dans le *JDM* que dans les autres

médias. L'exactitude la plus haute est d'ailleurs obtenue par le modèle peaufiné sur le *JDM* et évalué sur le *JDM* (0,988). L'exactitude la plus basse est obtenue par le modèle peaufiné sur *Le Soir* et évalué sur *La Presse* (0,752). L'ensemble de test de *La Presse* est le plus difficile à classer pour tous les modèles (excepté celui peaufiné sur des articles de *La Presse*), ce qui sous-entend qu'à l'inverse du *JDM*, *La Presse* semble être le média où la distinction entre les articles d'opinion et d'information est la plus difficile à cerner au niveau du texte. A l'inverse, le modèle peaufiné sur *La Presse* se transfère plutôt bien aux autres médias. Une hypothèse pour expliquer le défi posé par *La Presse* aux autres modèles serait liée au positionnement éditorial de *La Presse*, média de référence au Québec, non lucratif et détenu par une fondation pour la liberté des journalistes. *La Presse* se démarque en effet par ses chroniques aux tournures narratives et d'investigation, qui pourraient compliquer la tâche aux modèles entraînés sur les autres médias.

Figure 1. Exactitude des 10 modèles CamemBERT sur les 10 ensembles de test.

Concernant l'exactitude des modèles multimédias, on peut constater que le modèle *Québec* montre naturellement une meilleure exactitude sur les ensembles de test québécois que sur les belges, et vice-versa pour le modèle peaufiné sur des articles des quatre médias belges. Le modèle *Belgique* obtient malgré tout une meilleure exactitude sur l'ensemble de test du *JDM* que sur tous les autres, consolidant la qualité d'exception du *JDM* dans ces résultats. Le modèle *Québec* affiche les résultats les plus élevés en moyenne sur les dix ensembles de test, ce qui montre que peaufiner un modèle *transformers* sur diverses sources de données permet d'augmenter encore plus sa généricité.

	Devoir	*Presse*	*JDM*	*Soleil*	*RTBF*	*Soir*	*Avenir*	*Libre*	*Québec*	*Belgique*
Médias québécois	93.7	93.3	91.2	89.5	87.4	86.4	81.7	88.6	94.1	88.0
Médias belges	89.9	88.1	91.8	88.9	90.7	88.0	87.0	92.3	90.4	94.7

Table 1. Exactitude moyenne (%) de chaque modèle CamemBERT (fine-tuné sur les 10 ensembles d'entraînement différents) évalué sur les 4 ensembles de test québécois et les 4 ensembles de test belges.

La Table 1 montre l'exactitude moyenne obtenue par les différents modèles CamemBERT sur les quatre ensembles de test de chaque pays. On observe que, à part celui peaufiné sur le *JDM*, tous les modèles atteignent en moyenne une meilleure performance sur les textes des médias

issus du même pays. Cette tendance ressort particulièrement lorsqu'on se focalise sur les modèles *Québec* et *Belgique*, suggérant qu'il existe une différence certaine entre les deux cultures journalistiques en termes de distinction entre presse d'*information* et d'*opinion*.

4.1.2. Modèles basés sur des traits

La Figure 2 montre les résultats obtenus en utilisant le modèle basé sur des traits, entraîné et évalué sur les mêmes dix ensembles d'entraînement et de test. Il apparaît d'abord que les résultats obtenus par les différentes versions du modèle basé sur des traits sont nettement moins élevés que ceux atteints par les modèles CamemBERT : les exactitudes oscillent ici entre 0,562 et 0,902 (de 0,752 à 0,988 pour CamemBERT), ce qui montre que CamemBERT possède une capacité de transfert de connaissances d'un média à l'autre plus élevée que le modèle statistique.

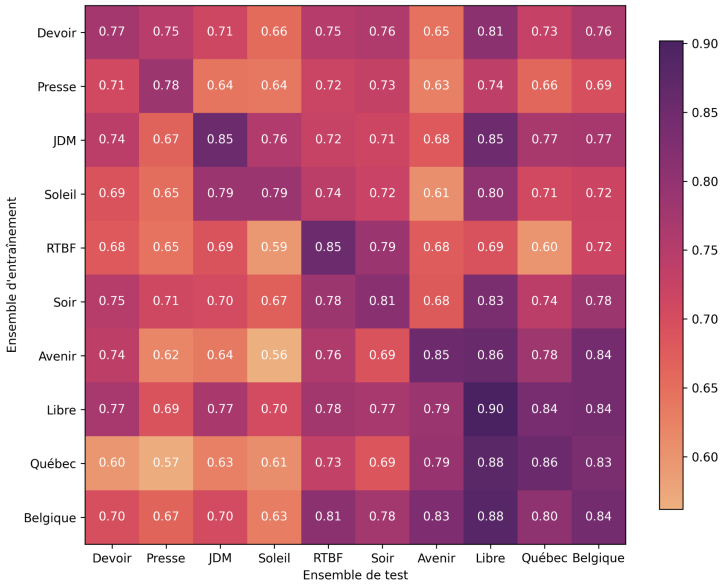

Figure 2. Exactitude des 10 modèles basés sur des traits sur les 10 ensembles de test.

Ensuite, il est possible d'observer que, comme pour CamemBERT, les modèles classent plus précisément les textes publiés par le même média que leur ensemble d'entraînement. La seule exception à cela concerne l'ensemble de test de *La Libre*, que plusieurs modèles classent avec plus de précision que leurs ensembles de test correspondants. Cette tendance ne s'observe pas dans la Figure 1, et à l'inverse l'ensemble de test du *JDM* n'apparaît pas comme aussi simple à classer pour les différents modèles statistiques que pour CamemBERT. De plus, la spécificité de *La Presse* montrée pour CamemBERT ne semble pas être capturée de même par les modèles statistiques, bien que les textes de *La Presse* restent difficiles à classer. Enfin, la puissance de généralisation des modèles *Québec* et *Belgique* n'est pas du tout aussi claire sur la Figure 2.

	Devoir	*Presse*	*JDM*	*Soleil*	*RTBF*	*Soir*	*Avenir*	*Libre*	*Québec*	*Belgique*
Médias québécois	72.3	69.3	75.4	72.9	65.2	70.8	63.8	73.1	59.9	67.4
Médias belges	74.1	70.3	74.2	71.7	75.3	77.8	78.9	80.9	77.2	82.4

Table 2. Exactitude moyenne (%) de chaque modèle basé sur des traits (entrainé sur les 10 ensembles d'entraînement différents) évalué sur les 4 ensembles de test québécois et les 4 ensembles de test belges.

Les moyennes présentées dans la Table 2 montrent qu'à l'exception du *JDM*, les modèles statistiques entraînés sur les différents médias québécois ne sont pas plus exacts sur les ensembles de test québécois que sur les ensembles belges, au contraire. Les modèles belges, quant à eux, sont significativement plus exacts sur les textes des médias belges que sur ceux des médias québécois. Ces résultats montrent que, pour les modèles statistiques, les textes de médias québécois sont globalement plus difficiles à classer, suggérant à nouveau une certaine différence entre les deux cultures journalistiques. Cette inégalité pourrait être due au fait que les dix-neuf premiers traits ont été sélectionnés à partir d'une étude réalisée sur un corpus d'articles belges de la *RTBF* (voir Section 3.2.2), bien que cette hypothèse soit mise à mal par le fait que le modèle entraîné sur la *RTBF* soit le moins efficace parmi les modèles belges.

4.2. Analyse des traits linguistiques prépondérants

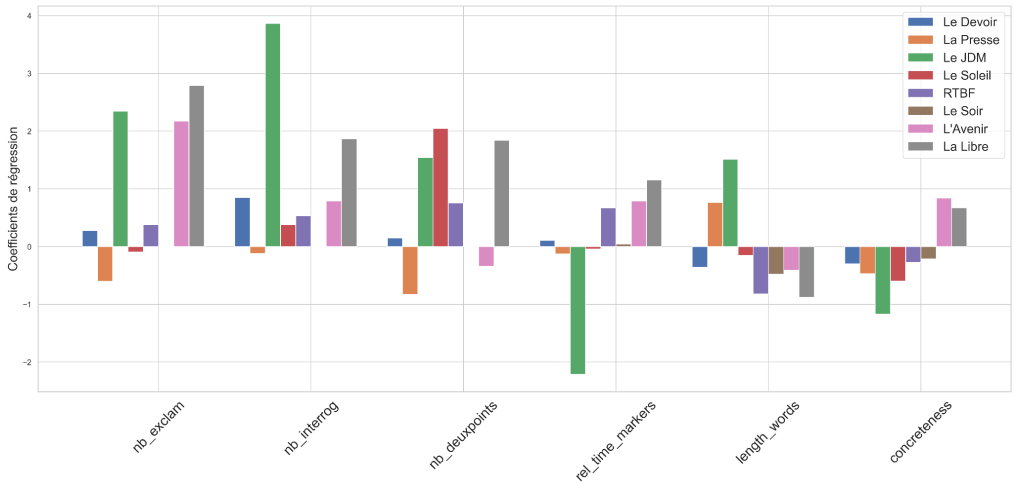

Figure 3. Poids des 6 traits avec le plus de variance entre les huit modèles.

Nous utilisons les coefficients de régression (poids) accordés à chaque trait par les différentes versions du modèle basé sur des traits afin d'interpréter plus en détails les résultats présentés dans la section 4.1. Les traits associés à des coefficients négatifs par un modèle sont ceux qui le portent à prédire la classe *information* pour un article, alors que les coefficients positifs le font tendre vers la classe *opinion*.

(1) *Notre société condamne-t-elle les discours haineux ? Cela dépend. Elle assimilera aux discours haineux la moindre critique des "minorités". Vous n'êtes pas convaincu que les drag queens ont leur place à l'école primaire ? Vous êtes haineux ! Vous ne croyez pas à la théorie du racisme systémique ? Toujours haineux !*

(extrait d'un article *opinion* publié le 3 avril 2023 par *Le Journal de Montréal*)

La Figure 3 visualise les six traits dont les poids présentent le plus de variance (ou le plus grand écart-type) d'un modèle à l'autre : points d'exclamation, points d'interrogation, deux-points, marqueurs de temps relatif, longueur moyenne des mots, et concrétude moyenne des mots. On peut observer que les poids du *JDM* pour certains traits apparaissent comme des valeurs particulièrement élevées ou basses, ce qui reflète son statut d'exception dans les résultats présentés dans la section 4.1. L'extrait (1) est un article d'opinion publié dans le *JDM* et qui est un des articles avec la plus haute fréquence de points d'exclamation et de points d'interrogation dans tous nos ensembles de test. Au Québec, le *JDM* est réputé pour ses éditoriaux "chocs" et polarisants, ce qui pourrait expliquer la distinction plus claire entre les articles d'*information* et

d'*opinion* observée dans les poids du modèle basé sur des traits, ainsi que la facilité qu'ont les différents modèles CamemBERT à classer les articles du *JDM*.

La Figure 3 montre aussi que le modèle entraîné sur *La Presse* attribue à divers traits des poids de signes opposés à ceux des autres modèles, ce qui pourrait expliquer la difficulté que présentent les différents modèles à classer ses textes. L'extrait (2) est issu d'un article de la classe *information* de *La Presse*, qui compte là aussi une des plus hautes fréquences de points d'exclamation, un signe de ponctuation pourtant typique du journalisme d'opinion. Cet extrait est caractéristique du style expressif et narratif de certains journalistes de *La Presse*. Il est également intéressant d'observer qu'on peut distinguer sur la Figure 3 une certaine divergence entre les quatre modèles entraînés sur des médias québécois et ceux entraînés sur des médias belges, notamment concernant les poids accordés aux marqueurs de temps relatifs et à la longueur moyenne des mots.

(2) *Dans Saint-Henri, un grand atelier silencieux continue d'être le refuge créateur du peintre Claude Tousignant, âgé de 88 ans. [...] Un capharnaüm dans lequel sa fille Zoé aimerait bien mettre son nez ! Pour mettre un peu d'ordre et nettoyer le plancher ! Le peintre n'est pas dérangé par cette apparence de désordre. C'est son antre !*

(extrait d'un article *information* publié le 18 août 2021 par *La Presse*)

Ensuite, la Figure 4 montre les huit traits dont les poids présentent le plus de variance entre les modèles statistiques *Québec* et *Belgique*. Pour les cinq premiers traits (points d'exclamation, conjonctions complexes, deux-points, points de suspension, mots de négation), il apparaît que le modèle *Québec* possède des poids particulièrement plus élevés que le modèle *Belgique*. Cela pourrait suggérer que le style d'écriture des articles d'opinion est plus affirmé dans les médias québécois que dans les médias belges. De plus, les deux modèles s'opposent dans l'orientation (positive ou négative) des poids qu'ils accordent aux marqueurs de temps relatifs et à la longueur moyenne des mots.

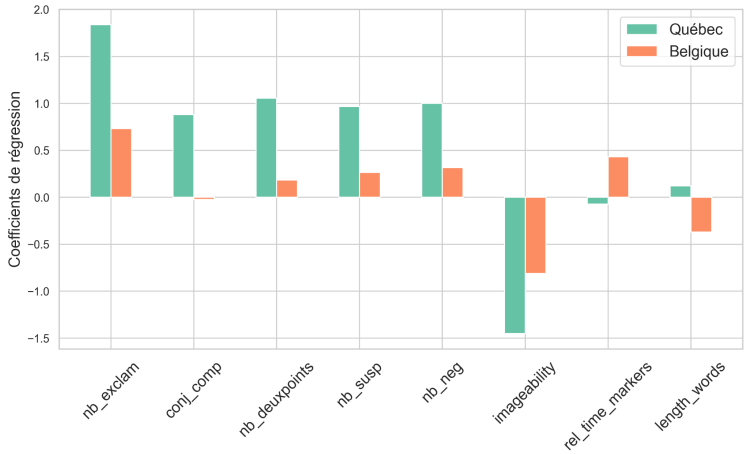

Figure 4. Poids des 8 traits avec le plus de variance entre les modèles Québec et Belgique.

5. Conclusion

Les résultats de nos expériences ont montré que, pour une tâche spécifique de classification d'articles d'opinion et d'information en français, un grand modèle neuronal *transformer* (CamemBERT) fait preuve d'une capacité de transfert de connaissances (vers des textes publiés par des médias différents de celui sur lequel il a été peaufiné) bien plus élevée qu'un modèle d'apprentissage automatique traditionnel (une régression logistique basée sur des traits

linguistiques). En analysant les coefficients de régression des modèles, nous avons également montré que les médias québécois et belges diffèrent quant à la manière dont les genres de l'opinion et de l'information sont caractérisés. Cette étude ouvre des pistes de recherche intéressantes. En particulier, pour mieux comprendre ce qui oppose ou rassemble ces deux cultures journalistiques, il serait pertinent de mener des analyses plus qualitatives, par exemple à l'aide d'entrevues avec des journalistes écrivant dans les différents médias étudiés. Il serait également intéressant d'inclure dans ces comparaisons des articles publiés par d'autres médias québécois et belges, voire issus d'autres pays francophones.

Bibliographie

Bogaert, De Marneffe M.-C., Descampe A., Escouflaire L., Fairon C. and Standaert F-X. (2024, sous presse). Sensibilité des explications à l'aléa des grands modèles de langage : le cas de la classification de textes journalistiques. *Traitement Automatique des Langues*, 64 (2).

Bonin P., Méot A., and Bugaiska A. (2018). Concreteness norms for 1,659 French words: Relationships with other variables and word recognition times. *Behavior research methods*, 50, 2366-2387.

Chalaby J.K. (1996). Journalism as an Anglo-American invention: a comparison of the development of French and Anglo-American journalism. *European journal of communication*, 11 (3), 303-326.

Charaudeau P. (2006). Discours journalistique et positionnements énonciatifs. *Semen*, 22.

Desrochers A., and Thompson G.L. (2009). Subjective frequency and imageability ratings for 3,600 French nouns. *Behavior research methods*, 41 (2), 546-557.

Devlin J., Chang M.W., Lee K. and Toutanova K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv:1810.04805.*

Escouflaire L. (2022). Identification des indicateurs linguistiques de la subjectivité les plus efficaces pour la classification d'articles de presse en français. In *Actes de la 29e Conférence sur le Traitement Automatique des Langues Naturelles*, 69-82.

Esser F. and Umbricht, A. (2012). Competing Models of Journalism? A Content Analysis of British, US American, German, Swiss, Italian and French Newspapers across Time. In *Conference of the International Communication Association in Phoenix*.

Grosse E.U. (2001). Évolution et typologie des genres journalistiques. Essai d'une vue d'ensemble. *Semen. Revue de sémio-linguistique des textes et discours*, (13).

Hallin D.C. and Mancini P. (2012). Comparing media systems between Eastern and Western Europe. In *Media transformations in the post-communist world: Eastern Europe's tortured path*, 15-32.

Liu Y., Ott M., Goyal N., Du J., Joshi M., Chen D., Levy O., Lewis, M., Zettlemoyer L. and Stoyanov V. (2019). RoBERTa: A robustly optimized BERT pretraining approach.

Lyu Q., Apidianaki, M., & Callison-Burch, C. (2024). Towards faithful model explanation in NLP: A survey. *Computational Linguistics*, 1-70.

Martin L., Muller B., Ortiz Suárez P.J., Dupont Y., Romary L., de la Clergerie É.V., Seddah D. and Sagot B. (2019). CamemBERT: a tasty French language model. *arXiv:1911.03894*.

Muñoz-Torres J.R. (2012). Truth and objectivity in journalism: Anatomy of an endless misunderstanding. *Journalism Studies*, 13 (4), 566-582.

Qasim R., Bangyal W.H., Alqarni M.A. and Ali Almazroi A. (2022). A fine-tuned BERT-based transfer learning approach for text classification. *Journal of healthcare engineering*.

Ravi K. and Ravi V. (2015). A survey on opinion mining and sentiment analysis: tasks, approaches and applications. *Knowledge-based systems*, 89, 14-46.

Schudson M. (2001). The objectivity norm in American journalism. *Journalism*, 2 (2), 149-170.

Tang H., Tan S. and Cheng X. (2009). A survey on sentiment detection of reviews. Expert Systems with Applications, 36 (7), 10760-10773.

Williams K. (2006). Competing models of journalism? Anglo-American and European reporting in the information age. *Journalistica*, 2.