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

COMFO : Corpus Multilingue pour la Fouille d'Opinions

Lamine FATY[1], Khadim DRAME[2], Edouard Ngor SARR[3,]
Marie NDIAYE[4], Yoro DIA[5] and Ousmane SALL[6]
[1,2,3,4,5]Université Assane Seck de Ziguinchor, SENEGAL
[6]Université Iba Der Thiam, SENEGAL
[7]Université Virtuelle du Sénégal, SENEGAL
{lamine.faty, khadim.drame, edouard-ngor.sarr, marie.ndiaye}@univ-zig.sn
yorodia2015@gmail.com
ousmane1.sall@uvs.edu.sn

RESUME

L'utilisation d'algorithmes de Machine Learning (ML) en fouille d'opinions notamment ceux d'apprentissage supervisé nécessite un corpus annoté pour entrainer le modèle de classification afin de prédire des résultats proches de la réalité. Malheureusement, il n'existe pas encore de ressources pour le traitement automatique de données textuelles exprimées dans le langage urbain sénégalais.
L'objectif de cet article est de construire un corpus multilingue pour la fouille d'opinions (COMFO). Le processus de constitution du corpus COMFO est composé de trois étapes à savoir la présentation de la source de données, la collecte et préparation de données, et l'annotation par approche lexicale. La particularité de COMFO réside dans l'intégration des langues étrangères (française et anglaises) et celles locales notamment le wolof urbain afin de refléter l'opinion collective des lecteurs sénégalais.

ABSTRACT

COMFO: Multilingual Corpus for Opinion Mining

The use of Machine Learning (ML) algorithms in opinion mining, particularly supervised learning algorithms, requires an annotated corpus to train the classification model in order to predict results that are close to reality. Unfortunately, there are still no resources for the automatic processing of textual data expressed in the Senegalese urban language.

The objective of this paper is to build a multilingual corpus for opinion mining (COMFO). The process of building the COMFO corpus is composed of three steps: presentation of the data source, data collection and preparation, and annotation by lexical approach. The particularity of COMFO lies in the integration of foreign languages (French and English) and local languages, notably urban Wolof, in order to reflect the collective opinion of Senegalese readers.

MOTS-CLES : Fouille d'opinions, commentaire en ligne, constitution de corpus, COMFO
KEYWORDS: Opinion Mining, Online Comment, Corpus Building, COMFO

Volume 1 : conférence principale, pages 297–304.
Cette œuvre est mise à disposition sous licence Attribution 4.0 International.

1 Introduction

Avec l'avènement des technologies web 2.0, nous assistons à un foisonnement de commentaires issus d'échanges à travers les réseaux sociaux, les sites de commerce, les sites dédiés à l'information, etc. Ces commentaires en ligne constituent d'immenses opportunités et sont souvent exploitées pour déterminer l'opinion majoritaire des intervenants. A cet effet, la Fouille d'Opinions (FO) [1][2] consiste à classer les documents (commentaires) en fonction des polarités positives, négatives et neutres à l'égard d'une entité ciblée (produit, service, phénomène, évènement, article journalistique, etc.). Aujourd'hui, nous assistons à un grand engouement autour de cette technique émergente avec l'utilisation d'algorithmes de Machine Learning (ML). L'application des algorithmes d'apprentissage supervisé en fouille d'opinions nécessite un corpus annoté pour entrainer le modèle de classification afin de prédire des résultats proches de la réalité.

Le travail que nous menons dans cet article trouve son contexte d'application dans le domaine de la presse en ligne sénégalaise. Il s'agit de l'exploration des commentaires issus de la presse en ligne sénégalaise. En général, la structure d'un commentaire en ligne est moins organisée. L'utilisation de ponctuations est beaucoup plus présente, les fautes d'orthographes sont très courantes. Tout ceci complique le travail de nettoyage préalable, surtout lorsqu'il est question de fouille d'opinions. En plus, les commentaires issus de ces sources sont écrits dans un langage libre qu'on appelle le langage urbain sénégalais. Malheureusement, peu de ressources sont disponibles pour le traitement automatique de données textuelles exprimées en langues nationales.

Notre objectif est de mettre en place et à disposition un corpus d'opinions étiqueté selon les classes suivantes : positive, négative ou neutre. Notre réelle motivation est de faciliter l'utilisation des algorithmes de Machine Learning (ML) pour la fouille d'opinions des commentaires sénégalais. L'innovation majeure de cette ressource bâtie COMFO réside dans l'intégration du langage urbain sénégalais afin d'interpréter convenablement les avis des utilisateurs. Dans un premier temps, nous allons faire l'état de l'art des travaux existants en insistant sur leurs limites. Ensuite, nous allons décrire la méthodologie d'annotation du corpus COMFO et l'évaluation des experts. En conclusion, nous allons faire le bilan et annoncer des perspectives.

2 Travaux connexes

Avec la nature multilingue des données issues des médias sociaux, beaucoup de travaux effectués récemment en FO (ou analyse de sentiments) intègrent plusieurs langues formelles et/ou informelles. Proksch et al. [4] ont présenté une approche multilingue d'analyse de sentiments pour estimer le conflit législatif dans la plupart des parlements européens en général. La construction du corpus est basée sur les débats parlementaires européens qui sont traduits automatiquement à l'aide du dictionnaire de Google. Ce dernier offre une base raisonnable pour l'analyse des sentiments dans différentes langues. Grljević et al. [5] ont présenté le premier corpus en langue serbe annoté manuellement pour les avis dans le domaine de l'enseignement supérieur. Les analyses statistiques et linguistiques du corpus ont révélé des informations utiles pour l'élaboration de règles manuelles d'analyse des sentiments. Hardalov et al. [6] ont proposé une méthode d'analyse de sentiments sur

un ensemble de données multilingues en provenance de plusieurs sources à l'égard d'une cible. Dans ce papier, les auteurs ont présenté les résultats d'une étude complète menée sur 15 jeux de données différents dans 12 langues de 6 familles de langues.

Bien que des efforts aient été faits pour l'analyse multilingue de sentiments basée sur une gamme de langues informelles, aucune ressource significative n'a été construite pour plupart des langues locales [7]. Les commentaires issus de la presse en ligne sénégalaise sont écrits dans le langage urbain sénégalais. Ce langage urbain sénégalais, d'une part, comporte des expressions issues de plusieurs langues (étrangères et nationales) et, d'autre part, modifie les caractéristiques orthographiques, voire grammaticales d'une langue afin de réduire sa longueur [8]. C'est dans ce contexte que nous nous plaçons pour proposer COMFO afin d'essayer de refléter l'opinion collective des lecteurs sénégalais. Les travaux connexes ci-dessus nous servirons de sources d'inspiration. A la suite de cette section, nous entamerons la méthodologie de construction de notre corpus.

3 Constitution du corpus COMFO

3.1 Présentation de la source données

Aujourd'hui, plusieurs sites dédiés à l'information ont vu le jour. Parmi ces multitudes de presse en lignes au Sénégal, Seneweb[1] jouit une position dominante. Ce site d'informations est considéré comme l'une des sources préférées des internautes sénégalais et de la diaspora. La figure 2 fourni des éléments statistiques sur lesquels nous nous sommes basés pour effectuer ce classement.

Libelle	Publications	Audiences	Partage	Popularité
www.seneweb.com	10246	82979014	3154	99,6194178
www.dakaractu.com	2007	200700	1241	0,24480769
www.senego.net	385	38500	3110	0,05040843
www.leral.net	308	30800	1135	0,03870268
Autres sites d'informations	12642	20200	6033	0,04666336
Total	**25588**	**83269214**	**14673**	**100**

FIGURE 2 : Popularité de sites d'informations sénégalais

Par ses trafics, Seneweb impose sa suprématie sur les autres sites du Sénégal. L'enrichissement et la prolifération des données de Seneweb ont rendu cette source utile et attrayante. Apres la présentation de sources de données, nous allons mettre en exergue notre stratégie de collecte.

3.2 Collecte et préparation de données

Notre système de collecte est basé sur OpinionScraper [9]. OpinionScraper est un scraper de collecte, de fusion et de formatage de données journalistiques. Il permet d'extraire des informations

[1] http://www.seneweb.com/

à partir de pages web de manière optimale et les formate en fonction des attributs notamment : *idCommentaire, texteCommentaire, auteurCommentaire, dateCommentaire, etc.* L'intérêt de l'utilisation de ce scraper est de constituer une base de données au format json facilement exploitable. La collecte des données est confrontée à plusieurs formes de bruits. Le bruit altère les données collectées et risque de rendre difficile l'apprentissage de la relation que l'on cherche à prédire, voire de rendre la modélisation impossible.

La préparation de données consiste à nettoyer ces bruits qui sont souvent de commentaires de publicités, de répétitions inutiles ou encore de commentaires composés d'une phrase incomplète. A cet effet, nous avons utilisé les expressions régulières. Les expressions régulières permettent de définir plusieurs critères de recherche en même temps afin d'identifier des motifs (patterns) à l'intérieur de contenu textuel. Sa puissance réside dans le fait qu'il s'applique indépendamment de la structure du document à analyser.

En définitive, nous disposons 13.500 commentaires dont 60% de commentaires en français (avec beaucoup d'expressions du français de la rue ou populaire) 37% en wolof urbain et 3% pour les langues. La figure 1 est une illustration de phrases construites à partir de mots ou groupes de mots issus de plusieurs langues notamment en français, anglais et wolof.

N°	Commentaire	Langue
1	Gnoune sounou khalè kamnani daal dièka nanou tè gawougnou magate machalla ndakh pt noire bii gawoul magate	Wolof
2	Mon pot les goûts et les couleurs ne se discutent pas- KOUNEK AK LILA DOYE – L'ESSENTIEL QUE VS VS AIMER LOLOU MON ÂME SOLO	Wolof-Française
3	thanks never give up we we support you for what you doing we appreciate you	Anglaise
4	Félicitations aux lions et nos encouragements aux béninois. Vive le Sénégal.	Française

FIGURE 1 : Extrait de commentaires sénégalais

3.3 Annotation par approche lexicale

L'annotation par l'approche lexicale consiste d'une part, à déduire l'opinion dégagée par un terme à l'aide d'un dictionnaire ou un lexique d'opinions et d'autre part à déterminer la polarité d'un commentaire à travers le calcul du score.

Nous avons utilisé SenOpinion [10] qui est un lexique d'opinions mis en place pour pallier à l'absence de ressources de fouille d'opinions dans le contexte sénégalais. C'est un lexique composé de mots, de locutions et d'expressions en français ou en wolof. Ce lexique est exclusivement destiné à étiqueter les commentaires écrits en langage urbain sénégalais. Pour trouver des correspondances entre les mots issus de notre base d'analyse (liste de termes à l'entrée) et ceux de notre lexique sans identificateurs uniques, nous préférons comparer des chaînes de lettres afin d'associer des étiquettes à chaque terme. Cette description peut être traduite en langage machine afin de permettre à l'ordinateur de procéder à l'étiquetage automatique d'opinions.

Le calcul s'effectue sur la base du score mesuré en fonction de la présence de termes issus des documents dans COMFO. Pour cela, on s'intéresse au résultat issu de ce calcul classique (voir Figure 3) :

- Soit $C = \{t_1, t_2, \dots, t_n\}$, un commentaire composé de n termes $t_1, t_2, \dots, t_n$;
- Soit P (Polarité), la valeur de chaque terme qui peut être -1 ou 1.

Le score d'un commentaire *C*, noté *Score(C)* est par définition la somme des polarités de termes qui composent le commentaire

$$Score(C) = \sum_{i=1}^{n} P(t_i)$$

- Si Score(C) > 0 alors *C* a une orientation positive ;
- Si Score(C) < 0 alors *C* a une orientation négative ;
- Si Score(C) = 0 alors *C* n'a pas une orientation (neutre).

FIGURE 3 : Classification de documents par approche lexicale

L'approche lexicale a surtout l'avantage de permettre des calculs rapides sur de grands corpus [11]. Cependant, la mise en place d'une ressource adaptée aux besoins d'applications spécifiques nécessite une évaluation par des experts.

4 Evaluation du corpus COMFO

4.1 Evaluation des experts

Le rôle des experts est déterminant dans le processus de construction d'une ressource linguistique telle qu'un corpus d'entrainement. Leur travail consiste à vérifier, à corriger si c'est nécessaire et à valider l'annotation proposée. Les experts sont composés de linguistes et de data scientistes. Nous avons, d'une part, les data scientistes de la formation en Intelligence Artificielle de l'Université

Virtuelle du Sénégal (promotion 1) et d'autre part les linguistes du master Lettre Moderne de l'Université Assane Seck de Ziguinchor. Les experts attribuent des polarités aux commentaires.

La validation des résultats issus de l'annotation manuelle est basée sur le système de vote majoritaire. Cette méthode de vote détermine la polarité finale d'un commentaire à l'aide du calcul de la médiane. En guise d'illustration, nous pouvons voir les polarités des documents de façon plus distincte à la figure ci-dessous (voir Figure 4).

Comments	Resources Tagging	Expert Evaluation
thanks never give up we we support you for what you doing we appreciate you	Neutral	Positive
Bro bayil senegalais yi ngay fowé koi mém si les lobby gordjigène te finance assume on sait tout c'est pas kom ca qu'on obtient un disc d'Or en plus ta music né pas riche et tu fau beaucoup de bruit	Positive	Negative
Ay way ma dieureum sey none.sou yalla ne waw koune dett si dett gay dee Wally reck eupeuna..	Positive	Positive
Nianthio diatawoulin allah barka que dieu vs bénissent	Neutral	Positive
Mashallah Maodo def nga Lou rafeet yaw sa xol dafa rafeet mashallah	Neutral	Positive
Salbe waha guoul si comba eume modou lo	Positive	Neutral
Il faut être sérieux ou même honnête et accepter votre défaite est claire devant les yeux de tout le monde	Positive	Negative
De toute façon quatre appuis ame na lithie desse mingui thie lokho C.N.G.	Positive	Neutral
Sois fairplay champion. Tu es tombé, il faut l, accepter. C,est le sport : on gagne, on perd ou c,est le nul. Lac 2 a gagné. Il faut te préparer et solliciter une revanche. COURAGE champion.	Negative	Positive
Lacs de guerre ne mérite pas ce victoires il n.a rien foutue lutteur de merde	Negative	Negative
Boy Niang tu es jeune nangoul dogal stpDanou ngua	Positive	Negative

FIGURE 4 : Extrait de commentaires annoté

Une fois les documents annotés, nous pouvons déterminer des statistiques avec les données à travers le niveau de commentaires. Ainsi, les statistiques au niveau des commentaires sont fournies à travers ces visualisations (voir figures 5 et 6).

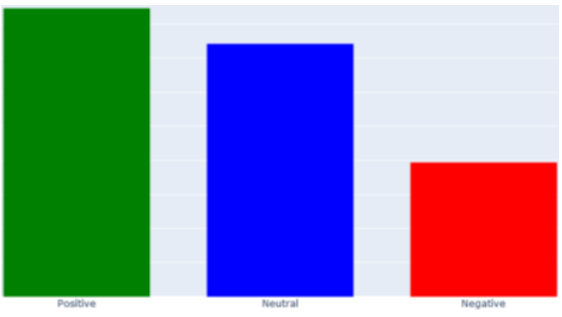

FIGURE 5 : Annotation lexicale

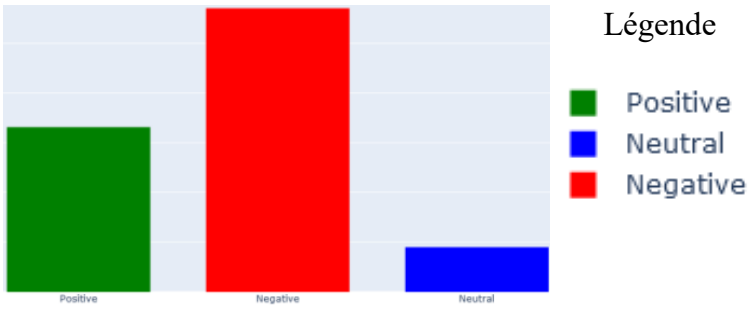

FIGURE 6 : Annotation lexicale

Cette visualisation permet une représentation synthétique et attrayante des résultats. La visualisation du corpus annoté à travers un histogramme nous permet de constater l'évolution des tendances en fonction des deux modes d'annotation. Lors de la discussion, nous fournirons des explications précises concernant les irrégularités ou inégalités dans l'évolution des tendances.

4.2 Discussion

Au total, nous avons annoté un jeu de données de 13.000 commentaires dont les statistiques sont fournies dans la figure 7.

Polarité	Resources Tagging	Expert Evaluation
Positive	42,7 %	33,4 %
Négative	19,9 %	57,5 %
Neutre	37,4 %	9,1 %

FIGURE 7 : Statistique des données annotées

Au regard de ces statistiques, nous constatons aisément une grande différence entre ces deux modes d'annotation. Cela s'explique par le fait que l'approche lexicale compare des chaînes de lettres afin d'associer des étiquettes à chaque terme. Une telle approche ignore les négations. La problématique des négations est un défi qui demeure en fouille d'opinions même avec les corpus écrits en anglais [12]. La complexité des négations réside dans le fait qu'elle modifie la polarité des commentaires initialement prévue [8]. A cela s'ajoute la complexité des types de commentaires que nous disposons avec les expressions issues du langage urbain sénégalais.

5 Conclusion

En définitive, la construction d'un corpus de fouille d'opinions relève de beaucoup d'efforts humains surtout dans un contexte où les outils de traitement automatique de langages naturels sont quasi inexistants. C'est une activité plus spécifiquement linguistique. Cette activité de collecte et d'annotation de données invite les experts à vérifier et à valider les résultats proposés. C'est une phase qui est assez longue mais permet d'avoir de bien meilleures performances. Elle permet de disposer d'un corpus propre et exploitable par les modèles de ML.

Dans la suite de notre analyse, nous utiliserons l'annotation des experts. Cette annotation a intégré des négations et le langage urbain sénégalais. Ce corpus sera mis à la disposition de la communauté scientifique pour les besoins de validation des méthodes qui sont expérimentées sur ces types de données. A l'avenir, nous comptons étendre les expérimentations sur les algorithmes de ML pour valider notre outil.

En raison de la nature multilingue des données des médias sociaux, une analyse basée sur une seule langue officielle peut comporter le risque de ne pas saisir le sentiment général du contenu en ligne.

Des efforts sont en train d'être faits pour la fouille d'opinions ou l'analyse de sentiments dans un contexte multilingue sur une gamme de langues informelles.

Références

[1] A. Jeyapriya et C. K. Selvi, « Extracting aspects and mining opinions in product reviews using supervised learning algorithm », in *Electronics and Communication Systems (ICECS), 2015 2nd International Conference on*, 2015, p. 548-552.

[2] B. Liu et L. Zhang, « A survey of opinion mining and sentiment analysis », in *Mining text data*, Springer, 2012, p. 415-463.

[3] M. Rushdi-Saleh, M. T. Martín-Valdivia, L. A. U. Lopez, et J. M. Perea-Ortega, « Bilingual experiments with an arabic-english corpus for opinion mining », in *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*, 2011, p. 740-745.

[4] S.-O. Proksch, W. Lowe, J. Wäckerle, et S. Soroka, « Multilingual sentiment analysis: A new approach to measuring conflict in legislative speeches », *Legis. Stud. Q.*, 2018.

[5] O. Grljević, Z. Bošnjak, et A. Kovačević, « Opinion mining in higher education: a corpus-based approach », *Enterp. Inf. Syst.*, p. 1-26, 2020.

[6] M. Hardalov, A. Arora, P. Nakov, et I. Augenstein, « Few-Shot Cross-Lingual Stance Detection with Sentiment-Based Pre-Training », *ArXiv Prepr. ArXiv210906050*, 2021.

[7] S. L. Lo, E. Cambria, R. Chiong, et D. Cornforth, « Multilingual sentiment analysis: from formal to informal and scarce resource languages », *Artif. Intell. Rev.*, vol. 48, n° 4, p. 499-527, 2017.

[8] L. Faty, M. Ndiaye, I. Diop, et K. Drame, « The complexity of comments from Senegalese online presses face with opinion mining methods », in *2019 14th Iberian Conference on Information Systems and Technologies (CISTI)*, 2019, p. 1-6.

[9] L. Faty *et al.*, « Opinion Scraper: A News Comments Extraction Tool for Opinion Mining », in *2020 3rd International Conference on Big Data and Computational Intelligence (ICBDCI)*, 2020, p. 1-9.

[10] L. Faty *et al.*, « SenOpinion: A New Lexicon for Opinion Tagging in Senegalese News Comments », in *2020 15th Iberian Conference on Information Systems and Technologies (CISTI)*, 2020, p. 1-6.

[11] S. Sun, C. Luo, et J. Chen, « A review of natural language processing techniques for opinion mining systems », *Inf. Fusion*, vol. 36, p. 10-25, 2017.

[12] L. Zhang, « Analyse automatique d'opinion: problématique de l'intensité et de la négation pour l'application à un corpus journalistique », PhD Thesis, Université de Caen, 2012.