SentiQ: Une approche logique-probabiliste pour améliorer la qualité de l'analyse des sentiments

Wissam Mammar Kouadri[*,**], Salima Benbernou[*], Mourad Ouziri[*]
Themis Palpanas[*], Iheb Ben Amor[**]

[*]Université de Paris, France
nom.prenom@u-paris.fr
[**]IMBA Consulting, France
prenom.nom@imbaconsulting.com

1 Introduction

L'opinion exprimée dans les données issues des réseaux sociaux représente un facteur majeur dans le processus de prise de décision. Cependant, malgré les avancés de recherche dans le domaine de l'analyse du sentiment, elle reste une tâche difficile en raison de la richesse et la complexité du langage naturel qui permet d'exprimer le même sentiment de différentes manières. Pour illustrer cette complexité, considérons les deux phrases : (a) Donald Trump softens tone on Chinese investments et (b) Trump drops new restrictions on China investment. Bien que (a) et (b) soient structurées différemment, elles sont sémantiquement équivalentes et expriment la même idée. Les travaux de recherche ont consenti que les textes sémantiquement équivalents doivent avoir la même polarité (Positive, Négative, ou Neutre). Cependant, nous avons constaté à travers des expérimentations intensives, que les algorithmes d'analyse de sentiments ne détectent pas la similarité entre les documents et extraient des polarités différentes conduisant à des incohérences *intra et inter-algorithmes*. Les incohérences intra-algorithme se produisent lorsqu'un algorithme extrait des polarités différentes de documents sémantiquement équivalents. Les incohérences inter-algorithmes représentent le cas où deux algorithmes extraient des polarités différentes d'un document. Dans ce papier, nous présentons SentiQ, un framework basé sur les réseaux logiques probabilistes de Markov (MLN) qui identifie et résout les deux types d'incohérences tout en améliorant la précision des algorithmes.

2 SentiQ : Une approche logique probabiliste pour une analyse de sentiment de qualité

Le vote majoritaire représente la solution triviale pour les problèmes d'incohérences intra et inter-algorithmes. Cependant, cette méthode attribut les mêmes pondérations à tous les algorithmes, ce qui implique une dégradation de la précision des algorithmes comme illustré dans Kouadri et al. (2020). Les limites du vote majoritaire ont été surmontées par des travaux de recherche comme celui de Ratner et al. (2020) qui attribue des pondérations aux algorithmes

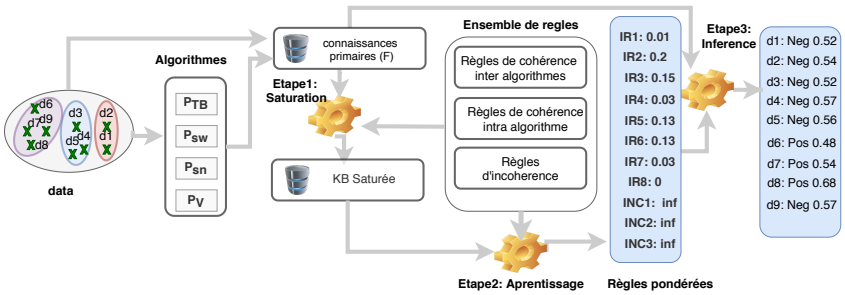

FIG. 1 – *Le framework SentiQ*

statistiquement en résolvant les incohérence inter-algorithmes, et celui de Ding et Riloff (2018) qui prend on considération la sémantique et améliore la précision en résolvant les incohérences intra-algorithmes. Néanmoins, aucun travail n'a traité les deux types d'incohérence en même temps. De ce fait, nous avons proposé SentiQ (FIG. 1), un framework basé sur MLN, qui résout les deux types d'incohérence et améliore la précision (Kouadri et al. (2020)).

•**Étape 1 - Le pré-traitement :** C'est la construction d'une base de connaissance $KB =< F, T >$, contenant les faits F et les règles T. Initialement, F contient les faits primaires, puis F sera saturé en appliquant deux types de règles d'inférences : (1) Les règles de qualité, dites *soft*, présument que les deux types de cohérence intra et inter-algorithmes soient respectées. (2) Les règles d'exploration d'incohérences.

•**Étape 2 - L'apprentissage de poids :** C'est l' attribution des pondérations aux algorithmes en se basant sur leurs cohérences : moins l'algorithme est cohérent, plus sa pondération est petite ; $min - \log P(Y = y|X = x) = min_{w_i} \log Z_x - \sum_i w_i n_i(x, y)$ tels que, X est l'ensemble F incohérents et saturés, Y l'ensemble des requêtes, Z_x un facteur de normalisation, et $n_i(x, y)$ le nombre de fois où les règles de qualité ne sont pas violées.

•**Étape 3 - L'inférence de la polarité adéquate :** C'est l'inférence des polarités les plus adéquates pour les documents. Cette polarité minimise les incohérences en prenant en considération la qualité des algorithmes et qui est représentée par leurs pondérations.

Les résultats obtenus sont prometteurs, montrent l'efficacité de SentiQ, et l'intérêt de résoudre les deux types d'incohérence pour améliorer la précision des algorithmes.

Références

Ding, H. et E. Riloff (2018). Weakly supervised induction of affective events by optimizing semantic consistency. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Kouadri, W. M., S. Benbernou, M. Ouziri, T. Palpanas, et I. B. Amor (2020). Sentiq : A probabilistic logic approach to enhance sentiment analysis tool quality. In *WISDOM@KDD 2020 : The 9th KDD Workshop on Issues of Sentiment Discovery and Opinion Mining*.

Ratner, A., S. H. Bach, H. R. Ehrenberg, J. A. Fries, S. Wu, et C. Ré (2020). Snorkel : rapid training data creation with weak supervision. *VLDB J. 29*(2), 709–730.