PROPOSITION DE SUJET DE THESE

Intitulé : Integration of knowledge and linguistic annotations to improve online content detection

Référence : **TIS-DTIS-2025-32**
(à rappeler dans toute correspondance) IAD DTIS

Début de la thèse :	**Date limite de candidature :**

Mots clés

Social data analysis, ontology, formal knowledge, linguistic annotation, hybrid AI models

Profil et compétences recherchées

Research Master

Knowledge representation: ontologies, conceptual graphs, linguistic annotation, supervised and unsupervised machine learning

Communication skills

Good level of French and English

Programming skills: Python, OWL

Présentation du projet doctoral, contexte et objectif

Context and motivation

Detection of specific online content remains a persisting issue in social media analysis. This includes extremism, online hate or offensive language, which are topics of interest for regulation of social platforms. Being complex phenomena, automatic methods have to consider a variety of features to capture their nature [1, 2]. Typically, automatic approaches for content detection are based on natural language processing and rely upon both fine-tuning using a specialized dataset and development of learning models [3].

Making sense out of large collections of online data and creating potential coherent representations for specific goals (e.g. crisis analysis, threat detection) is a challenging task. Identifying existing resources to complete or complement new approaches also requires a considerable effort. The goal of the thesis is to investigate to what extent the joint integration of formal knowledge and linguistic annotations, with data-driven learning algorithms is able to improve the automatic detection of specific types of data such as extremism or online hate.

Objectives and research directions

The scientific objective of this thesis is to investigate the potential of integrating knowledge and linguistic annotations in order to improve the detection of specific types of online content. Main research directions are elaborated hereafter.

Emotional annotation for social data analysis

In the field of Natural Language Processing (NLP), automatic analysis of emotions in written texts is generally addressed exclusively through the notion of emotional category (e.g. \textit{Joy, Fear}, etc.) - often with a focus on a sole linguistic mean to express emotions, the emotional lexicon. However, as pointed out by linguistics and NLP very recent works [12] this is not sufficient to explore, and then to identify, emotions in their diversity of modes of expressions in texts. Moreover, from a strictly NLP and/or information extraction point of view, there is a need to consider the huge diversity of emotion expressions (thus, not only the strictly lexical ones) in order to better quantitatively capture the emotions. For example, emotions can be expressed by interjections, as described in [14] or emojis [15]. They can also be revealed by appraisals, behaviors or suggestions as presented in [13].

User engagement for social data analysis

The concept of user engagement is related to speakers' own perspective on informational content and indicates whether they assert with confidence what they perceive in the environment, know from their experience, interact with, or attend to. In some particular cases, for example, when indirectly reporting, the authors can only assume, to varying degrees of certainty, and this aspect is also covered by engagement.

Although engagement is a quite well established notion in linguistics and has been explored at the intersection of modality, evidentially, and commitment categories there has been less work on considering speaker/user engagement for social data analysis. However, previous research confirms that, for open-dialogue systems, taking into account user engagement as real-time feedback benefits the analysis of social interactions More specifically, online users use emojis to enrich the context and convey additional emotions, and using emojis increases user engagement.

Leveraging word embeddings with taxonomies and ontologies

Word embeddings is a generic name for a class of models using shallow supervision to capture semantics directly from data without additional background. Well-known word embedding algorithms include namely Word2Vec [5], LINE [6] and GloVe [7]. Those models build low-dimension vectorial representations of words, and this can be done without training a complete language learning model, with a lot of overhead and a very long training time.

Although achieving reasonable precision values, incorporating relevant features from semantic models such as WordNet [4] or Paraphrase Database (PPDB) [8] has been shown to improve the quality of word embeddings for specific tasks. Those models describe common knowledge; the goal of the thesis is to integrate specific knowledge from customized and proprietary ontologies. More specifically, we will investigate the following directions to build hybrid systems combing knowledge models and words embeddings:

Lately, some researchers extended the ides of graphs embeddings to ontologies and taxonomies, with different and explicitly named conceptual relations. Several research efforts addressed this last research direction to investigate how the use of semantic relationships such as synonymy, antonymy, hypernymy, extracted from several semantic models in order can improve the quality of word embeddings, which is to say the accuracy of the semantics captured by vectorial representation of words. The thesis will investigate how specific relations modeled by domain ontologies [9, 10, and 11] can be added to word embedding’s and captured by the resulting hybrid model. The application envisioned for this task is the detection of extremist content in streams of online data.

Research conducted is expected to investigate the integration of knowledge and linguistic features for the development of hybrid artificial intelligence methods, at the crossroads of knowledge representation and machine learning.

References:

1. Schäfer, J., & Kistner, E. (2023). HS-EMO: Analyzing Emotions in Hate Speech. In Proceedings of the 19th Conference on Natural Language Processing (KONVENS 2023) (pp. 165-173).

2. Yin, W., Agarwal, V., Jiang, A., Zubiaga, A., & Sastry, N. (2023). Annobert: Effectively representing multiple annotators’ label choices to improve hate speech detection. In Proceedings of the International AAAI Conference on Web and Social Media (Vol. 17, pp. 902-913).

3. Ljubešić, N., Mozetič, I., & Novak, P. K. (2023). Quantifying the impact of context on the quality of manual hate speech annotation. Natural Language Engineering, 29(6), 1481-1494

4. Miller, G. A. (1995). WordNet: a lexical database for English. *Communications of the ACM*, *38*(11), 39-41.

5. Church, K. W. (2017). Word2Vec. *Natural Language Engineering*, *23*(1), 155-162.
6. Tang, J., Qu, M., Wang, M., Zhang, M., Yan, J., & Mei, Q. (2015, May). Line: Large-scale information network embedding. In *Proceedings of the 24th international conference on world wide web* (pp. 1067-1077).
7. Pennington, J., Socher, R., & Manning, C. D. (2014, October). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1532-1543).
8. Pavlick, E., Rastogi, P., Ganitkevitch, J., Van Durme, B., & Callison-Burch, C. (2015, July). PPDB 2.0: Better paraphrase ranking, fine-grained entailment relations, word embeddings, and style classification. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)* (pp. 425-430).
9. Dragos, V., Battistelli, D., Etienne, A., & Constable, Y. (2022, June). Angry or Sad? Emotion Annotation for Extremist Content Characterization. In *13th Language Resources and Evaluation Conference*.10. Battistelli, D., Bruneau, C., & Dragos, V. (2020). Building a formal model for hate detection in French corpora. *Procedia Computer Science*, *176*, 2358-2365.]
11. Dragos, V., Battistelli, D., & Kelodjoue, E. (2018, July). Beyond sentiments and opinions: exploring social media with appraisal categories. In *2018 21st International Conference on Information Fusion (FUSION)* (pp. 1851-1858). IEEE
12. Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary,
Éric Villemonte de La Clergerie, Djamé Seddah, and Benoît Sagot. 2019. Camembert: a tasty french language model. arXiv preprint arXiv:1911.03894
13. Roman Klinger. 2023. Bridging emotion role labeling and appraisal-based emotion analysis. arXiv preprint arXiv:2309.02092.
14. Gustave Cortal, Alain Finkel, Patrick Paroubek, and Lina Ye. 2023. Emotion recognition based on psychological components in guided narratives for emotion regulation. arXiv preprint arXiv:2305.10446
15. Delphinel Battistelli, Valentina Dragos, and Jade Mekki. 2023b. Annotating social data with speaker/user engagement. illustration on online hate characterization in french. In International Conference on Computing and Communication Networks 2023: ICCCN 2023

Collaborations envisagées

Collaboration avec le laboratoire MODYCO : Modèles, Dynamiques, Corpus, Unité de Recherche mixte CNRS & Université Paris Nanterre

Laboratoire d'accueil à l'ONERA Département : Traitement de l'information et Systèmes Lieu (centre ONERA) : Palaiseau **Contact** : Valentina Dragos Tél. : 01 80 38 65 65 Email : valentina.dragos@onera.fr	**Directeur de thèse** Nom : Valentina Dragos Laboratoire : DTIS/MIDL Tél. : 01 80 38 65 65 Email : valentina.dragos@onera.fr

Pour plus d'informations : https://www.onera.fr/rejoindre-onera/la-formation-par-la-recherche