GeoNLPlify : Une augmentation spatiale de corpus liés aux crises pour des tâches de classification

Rémy Decoupes[*,***], Mathieu Roche[**,***], Maguelonne Teisseire[*,***]

[*] INRAE, F-34398 Montpellier, France
prenom.nom@inrae.fr,
[**] CIRAD, F-34398 Montpellier, France
prenom.nom@cirad.fr
[***] TETIS, Univ. Montpellier, AgroParisTech, CIRAD,
CNRS, INRAE, Montpellier 34090, France

Résumé. L'article *"Deux cygnes retrouvés mort au **Parc de la Tête d'Or** à **Lyon"*** parle-t-il de l'épidémie de grippe aviaire ? Nos travaux proposent d'utiliser l'information spatiale pour générer des données artificielles étiquetées afin d'améliorer les classifications de textes basées sur BERT. Ainsi, après avoir mis en évidence, par des méthodes d'explicabilité, l'importance de l'information spatiale dans les corpus liés à des crises, nous proposons différentes stratégies d'augmentation de données qui tirent profit de ce constat. Notre méthode, GeoNLPlify, est évaluée sur des jeux de données publics (PADI-web et CrisisNLP) et comparée aux augmentations de données classiques.

1 Introduction

La dégradation de l'environnement et les effets croissants du changement climatique provoquent une augmentation du nombre de catastrophes et de leurs impacts. Pour permettre une meilleure gestion de ces situations, il devient nécessaire de faire appel à des méthodes d'analyse de données performantes. Le problème auquel nous sommes alors confrontés est le peu de données disponibles. En effet, comme la rareté et la non-similitude des événements sont importants (Buntain et al., 2020), il devient peu pertinent d'appliquer des méthodes d'adaptation de domaine.

En parallèle, le développement des modèles de langue (Large Language Models (LLM)), basés sur les mécanismes d'attention (Vaswani et al., 2017), s'est accru ces dernières années avec des performances exceptionnelles. Même si les LLM sont destinés à être utilisés par transfert d'apprentissage sur des corpus plus petits, ils ont toujours besoin d'un ensemble de données assez important. Différentes techniques d'augmentation de données ont été développées en Traitement Automatique du Langage Naturel (TALN) (Bayer et al., 2022). Leurs objectifs sont d'améliorer les performances d'un modèle de classification de texte en générant artificiellement de nouvelles données étiquetées pour augmenter la taille du corpus d'apprentissage. Cependant, plusieurs approches sont inefficaces quand des LLM sont utilisés car ces derniers sont invariants à certaines transformations (Longpre et al., 2020) telles que le remplacement

de lettres ou de mots. L'annotation de données par des experts étant très coûteux, un défi est de trouver de nouvelles méthodes d'augmentation de données ayant un impact positif sur les classificateurs LLM.

Pour une tâche de classification de texte dans le domaine des gestions de crises, nous faisons l'hypothèse que les données associées possèdent une très forte composante spatiale qu'il faut utiliser. C'est pourquoi, nous proposons GeoNLPlify [1], une librarie python comportant trois méthodes d'augmentation de données fondées sur les informations spatiales contenues dans les textes. Nous montrons que GeoNLPlify a un impact significativement positif sur les performances des classificateurs LLM qui ont été entraînés sur des corpus pour lesquels la spatialité est importante. Pour souligner le rôle de l'information spatiale dans les corpus liés à des crises, nous proposons une analyse des techniques d'explication pour les modèles d'apprentissage profond appliqués aux modèles LLM. L'objectif est de détecter les catégories de mots qui ont le plus grand impact sur les prédictions du classifieur. L'importance de l'information spatiale, pour les corpus traitant de situations de crise, est démontrée et discutée dans cet article. Notre approche GeoNPlify est évaluée sur 2 corpus : PADI-Web [2] (Arsevska et al., 2018) et crisisNLP [3] (Imran et al., 2016) et comparée aux méthodes d'augmentation de données récentes en TALN (Ma, 2019). Les modèles entraînés à partir de tels corpus, pour lesquels la spatialité compte, voient leurs performances considérablement augmentées par GeoNLPlify.

Dans la suite de cet article, après un état de l'art présenté en section 2, nous détaillons notre méthode en section 3 pour ensuite décrire et discuter les différentes évaluations réalisées en section 4. Enfin, nous dressons le bilan et les perspectives en conclusion.

2 État de l'art

Le traitement automatique du langage naturel (TALN) a bénéficié de l'émergence des modèles de langues (LLM). Plusieurs d'entre eux ont été mis à disposition de la communauté tels que BERT (Devlin et al., 2019) ou RoBERTa (Liu et al., 2019). Ces modèles sont non spécifiques et peuvent, par un ré-entraînement, être spécialisés, ou affinés, à un domaine particulier comme la gestion de crises qui nous intéresse dans cette étude. Malheureusement, ce domaine souffre d'un manque de données (Buntain et al., 2020).

Pour surmonter ce problème, les jeux de données annotées doivent être *augmentés*. Une première solution consiste à demander à un expert d'étiqueter de nouvelles données mais ceci n'est pas toujours envisageable. Une solution alternative est de créer artificiellement de nouvelles données étiquetées. Il existent, notamment, des méthodes basées sur des heuristiques (par exemple étiqueter négativement un texte s'il comporte le mot « pleurs »). Cependant ce type de règles n'est pas toujours évident à définir. Une autre piste consiste à entraîner un deuxième modèle qui générera des données pseudo-étiquetées de confiance (Li et al., 2020). Cependant, deux limites peuvent être opposées à ce type d'approche : (i) la sur-représentation des données pour lesquelles la classification est simple et (ii) accroître le risque d'apporter des pseudo-étiquettes erronées.

Contrairement aux méthodes précédentes, l'augmentation de données (Data Augmentation DA) n'opère pas sur des données non étiquetées pour générer artificiellement des pseudo-

1. https ://github.com/remydecoupes/GeoNLPlify
2. https ://padi-web.cirad.fr/
3. https ://crisisnlp.qcri.org/lrec2016/ lrec2016.html

étiquettes. L'objectif est de faire quelques variations de données étiquetées afin d'en générer de nouvelles en garantissant leur qualité. Plusieurs stratégies de DA en TALN ont été fournies par la communauté. Ainsi, Easy Data Augmentation (Wei et Zou, 2019) vise à créer une variation du contenu des données étiquetées en remplaçant ou en subsistant ou en ajoutant des caractères ou des mots par synonymie ou par des mécanismes aléatoires. Un LLM pourra également être utilisé pour remplacer un mot par un autre (Kobayashi, 2018). Malheureusement, les classificateurs LLM, de par leur nature, peuvent être insensibles à ce genre de variations (Longpre et al., 2020). D'autres tâches du TALN, telles que la traduction ou la synthèse peuvent bénéficier d'une rétro-traduction (traduire une phrase dans une autre langue et revenir à l'originale) (Sennrich et al., 2016). La DA est également appliquée pour évaluer les modèles, appelé entraînement contradictoire, en introduisant certaines variations dans les données jusqu'à ce que les modèles infèrent une mauvaise étiquette (Morris et al., 2020).

Dans cet article, nous proposons une stratégie originale d'augmentation des données fondée sur l'information spatiale. Pour valider notre hypothèse, nous montrons que le classificateur ré-entraîné s'appuie sur des types de mot en lien avec la spatialité. Pour cela, nous utilisons des cartes de saillance « LIME » (Ribeiro et al., 2016).

3 L'approche GeoNLPlify

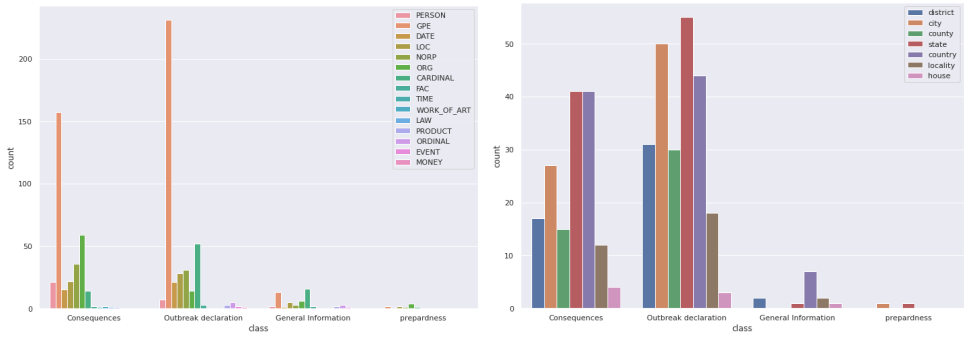
Pour améliorer les re-entraînements de LLM, nous proposons GeoNLPlify, un ensemble de nouvelles approches d'augmentation de données basé sur l'information spatiale contenue dans les textes. Comme indiqué par Longpre et al. (2020), les LLM peuvent être invariants à certaines augmentations de données. Ainsi, avant d'introduire GeoNLPlify, nous mettons en évidence, via une methode d'explicabilité, l'importance de l'information spatiale pour la classification de textes sur des corpus liés aux crises et pour lesquels la spatialité compte.

3.1 Carte des mots saillants

Après avoir ré-entraîné un modèle pré-entraîné, nous dressons, pour chaque document, des cartes de mots saillants. Les données utilisées pour l'entraînement sont issues de PADI-Web (Arsevska et al., 2018), un corpus annoté traitant de santé animale. Le modèle pré-entraîné RoBERTA (Liu et al., 2019) est utilisé car il obtient les meilleurs performances sur le corpus de base non augmenté.

Afin de comprendre sur quels types de mot le modèle RoBERTa ré-entraîné fonde ses prédictions, nous utilisons LIME (Ribeiro et al., 2016) comme méthode d'interprétation. Cette méthode fait des variations sur les mots d'un document ce qui permet d'entraîner un modèle explicatif permettant de reproduire le comportement de notre classificateur "boîte noire" LLM. Comme les méthodes d'explicabilité ou d'interprétabilité appliquées à des LLM peuvent être contradictoires, nous ne l'utilisons pas ici comme une explication complète de notre classificateur mais nous considérons le résultat LIME comme un moyen pour formuler une hypothèse que nous nous efforcerons de valider par la suite.

Aussi, pour exploiter les résultats LIME, nous extrayons les trois mots les plus saillants pour chaque document du corpus annoté. Pour chaque jeton saillant, nous appliquons la reconnaissance d'entité nommée (NER) qui classe le jeton dans des catégories prédéfinies telles que



(a) Distribution des mots les plus saillants selon les catégories NER

(b) Les différentes granularités spatiales des entités GPE

FIG. 1 – Analyse des types de mots saillant par phase de crise

les noms de personnes, les organisations, les lieux ou les entités géopolitiques (GPE) qui sont des villes/états/pays.

Le jeton GPE (localisation) contribue le plus, en nombre (illustré sur la figure 1a), aux prédictions locales fines des classificateurs RoBERTa et ceci pour toutes les classes. C'est pourquoi, nous nous appuyons sur la saillance de l'information spatiale pour proposer plusieurs stratégies d'augmentation des données, stratégies décrites dans la section suivante.

3.2 L'augmentation de données avec GeoNLPlify

GeoNLPlify est un ensemble de trois stratégies d'augmentation de données qui permet de réaliser des variations de données annotées en se concentrant sur l'information spatiale pour augmenter la taille du corpus d'apprentissage. En effet, comme souligné dans la section précédente, les jetons porteurs d'informations géographiques aident les classificateurs à faire leurs prédictions lorsqu'ils travaillent avec un ensemble de données lié à des crises. Notre intuition est que le niveau hiérarchique des informations spatiales a une influence sur le classement : être au bon niveau spatial permet de mieux rendre compte de la situation locale en temps de crise. Par exemple, une déclaration d'épidémie se concentrera au niveau de la ville ou de la région où les cas sont apparus, alors que les conséquences seront signalées au niveau du pays. Ceci est illustré, pour l'ensemble de notre corpus, par la figure 1b. Afin d'évaluer l'hypothèse posée, nous définissons trois stratégies.

Augmentation par généralisation spatiale :

Il s'agît de dupliquer les documents annotés qui contiennent des jetons GPE au niveau de la ville en les remplaçant par leurs pays. Par exemple, le titre de ce nouvel article : *"2 cas de virus Powassan confirmés dans le* ***New Jersey****"* sera dupliqué par *"2 cas de virus Powassan confirmés dans le* ***États-Unis****"*.

Augmentation par spécialisation spatiale :

Contrairement au principe d'augmentation précédent, l'objectif est ici de dupliquer les documents contenant le GPE au niveau d'un pays par une ville choisie au hasard (nous descendons dans la hiérarchie spatiale). Pour cela, nous utilisons la base de données "world cities" de

simplemaps [4]. Par exemple, nous dupliquons le document *"**Kenya** émet une alerte sur l'épidémie de fièvre aphteuse"* en créant *"**Jakarta** émet une alerte sur l'épidémie de fièvre aphteuse"*. Nairobi est choisie au hasard dans la liste des villes du Kenya.

Augmentation par synonymie ou équivalence spatiale :

Pour cette dernière proposition d'augmentation, nous faisons des variations de même niveau pour les jetons GPE. La variante est, encore une fois, choisie au hasard parmi les "villes du monde" de simplemaps. Par exemple, le document *"2 cas de virus Powassan confirmés à **Vancouver"*** sera dupliqué par *"2 cas de virus Powassan confirmés à **Glasgow"***

4 Expérimentation

4.1 Protocole expérimental

4.1.1 Description des données

L'analyse comparative est effectuée sur deux ensembles de données liés aux crises (PADI-Web (Arsevska et al., 2018) et CrisisNLP (Imran et al., 2016)). Parmi l'ensemble des articles de presse traitant de santé animale récoltés par **PADI-Web**, 300 ont été manuellement annotés par un expert. Au nombre de 5, ces labels correspondent à une des phases d'une crise. **CrisisNLP**, quant à lui, est ensemble de tweets collectés lors de 19 crises (entre 2013 et 2015). 11570 tweets ont été ensuite manuellement annotés parmi 14 types d'information utilisés en tant de crise. Ces deux corpus ont des répartitions très déséquilibrées entre leurs classes.

4.1.2 Augmentation des données

GeoNLPlify : Augmentation spatiale des données L'augmentation des données spatiales repose sur un processus en trois étapes : NER, géocodage et variation spatiale. Pour la tâche de NER, notre pipeline utilise un algorithme spaCy [5]. La deuxième étape se concentre sur les jetons identifiés comme des entités géographiques (GPE). À l'aide des données OpenStreetMap (OSM) [6] via le geocoder "photon" [7], le pipeline récupère les informations spatiales du jeton, telles que sa granularité spatiale (c.-à-d. ville/comté/état/pays).
Selon les méthodes d'augmentation des données spatiales, la troisième étape utilise les informations du niveau spatial pour créer une variation (généralisation, spécialisation ou synonymie spatiale).

Augmentation des données par des approches TALN Pour les approches classiques d'augmentation de données TALN, nous utilisons la bibliothèque python nlp_aug (Ma, 2019). Cette bibliothèque fournit plusieurs augmentations à différents niveaux (caractère, mot et phrase) à travers de multiples approches (incorporation contextuelle, synonyme, rétrotraduction, variation aléatoire, ...). Deux augmentations ont été utilisées pour la comparaison avec GeoNLPlify. Elles font, comme GeoNLPlify, des variations au niveau mot : synonyme (basé sur WordNet (Miller, 1995)) et incorporation contextuelle de mots (utilisant le modèle BERT). Les deux

4. https://simplemaps.com/data/world-cities
5. https://spacy.io/models/en#en_core_web_trf
6. https://www.openstreetmap.org
7. https://photon.komoot.io

	prepardness	Consequences	General information	Other	Outbreak declaration
none	0.4	0.54	0.6	0.55	0.83
geo_generalization	0.62	0.7	0.63	0.64	0.87
geo_specialization	0.85	0.88	0.83	0.79	0.96
geo_spatial_synonym	0.8	0.88	0.81	0.69	0.97
nlp_substitute	0.61	0.65	0.68	0.56	0.82
nlp_synonym	0.75	0.76	0.79	0.84	0.88

TAB. 1 – Comparaison des moyennes F1-score sur PADI-Web

	affected_people	caution_and_advice	deaths_reports	disease_signs	disease_transmission	evacuations	donation	damage	injured_dead	missing	irrelevant	other_information	prevention	emotional_support	treatment
none	0.82	0.58	0.63	0.66	0.72	0.75	0.82	0.75	0.9	0.66	0.75	0.69	0.69	0.84	0.81
geo_generalization	0.88	0.65	0.64	0.61	0.73	0.79	0.84	0.79	0.9	0.72	0.76	0.74	0.6	0.86	0.69
geo_specialization	0.91	0.81	0.8	0.65	0.78	0.92	0.93	0.89	0.96	0.86	0.83	0.87	0.62	0.92	0.77
geo_spatial_synonym	0.92	0.82	0.87	0.72	0.81	0.92	0.93	0.9	0.97	0.88	0.84	0.87	0.7	0.93	0.8
nlp_substitute	0.64	0.43	0.43	0.48	0.56	0.55	0.63	0.56	0.66	0.47	0.57	0.62	0.47	0.64	0.54
nlp_synonym	0.76	0.65	0.6	0.6	0.66	0.76	0.86	0.76	0.89	0.65	0.82	0.75	0.56	0.88	0.68

TAB. 2 – Comparaison des moyennes F1-score sur CrisisNLP

approches utilisent la même stratégie pour créer des variations. Elles sélectionnent au hasard des jetons dans le document, puis remplacent ces jetons par l'une de ses variantes (en moyenne 30% des jetons sont modifiés).

Les cinq méthodes d'augmentation de données sont appliquées aux deux corpus. Comme GeoNLPlify ne peut générer des variations de documents qu'à condition qu'ils possèdent une information spatiale, son augmentation n'est pas uniforme. Certaines classes peuvent sembler être désavantagées par la méthode GeoNLPlify alors que les méthodes NLP_aug augmentent, quant à elles, uniformément toutes les classes.

4.1.3 Entraînement et évaluation

Pour comparer les augmentations de données, un processus de validation croisée est utilisé en dix tours (plis). Les plis préservent le déséquilibre des classes. RoBERTa est ensuite spécialisé en 3 époques sur un serveur possédant une carte graphique NVIDIA V100 et 315 Go de RAM. L'entraînement est assuré par la bibliothèque python huggingFace [8]. Les métriques d'évaluation sont calculées sur le jeu de données initial sans augmentation. La moyenne de ces métriques (score F1, rappel et précision) des 10 plis est calculée pour chaque méthode d'augmentation de données.

4.2 Résultats

Après avoir ré-entraîné RoBERTa sur 10 tours, les moyennes des F1-score sont calculées pour chaque classe afin d'évaluer l'apport des différentes méthodes d'augmentation de données. Les tableaux 1 & 2 proposent ces résultats.

8. https ://huggingface.co/

Pour **PADI-Web**, la première constatation est que toutes les méthodes d'augmentation obtiennent de meilleurs résultats que le corpus non augmenté (portant le label "none" dans les tableaux). Deuxièmement, deux méthodes de GeoNLPlify, *Spatial Synonym* et *Specialization* obtiennent les meilleurs résultats hormis pour la classe "Ohter" pour laquelle *nlp_synonym* est meilleur. Le fait que cette classe ne possède que très peu d'informations géographiques n'explique pas, à lui tout seul, les moins bons résultats de GeoNLPlify puisque nous ne constatons pas cette baisse de performance pour la classe "General information", pauvre également en informations spatiales. La classe "other" est aussi trop bruitée (c'est à dire hors sujet)
En ce qui concerne **CrisisNLP**, *Spatial Synonym* et *Specialization* de GeoNLPlify obtiennent de nouveau les meilleures résultats, à part pour la classe "treatment" pour laquelle aucune DA n'améliore la classification.

5 Conclusion

Après avoir souligné l'importance des mots porteurs d'informations spatiales des jeux de données liés aux gestions de crises pour les modèles de classification de texte, nous avons proposé de nouvelles méthodes d'augmentation de données basées sur la spatialité. Cet ensemble de techniques, regroupées au sein du paquet python GeoNLPlify [9], est comparé aux méthodes traditionnelles d'augmentation de jeux de données. Ces comparaisons ont été menées sur deux corpus différents (PADI-Web et CrisisNLP). GeoNLPlify obtient les meilleurs performances pour la quasi totalité des classes. L'apport de GeoNLPLify pour spécialiser des modèles de type BERT pour des tâches de classification de texte traitant de crises est donc démontré.

Les travaux futurs se focaliseront sur les méthodes de combinaison de techniques d'augmentation de données afin d'améliorer, encore, la qualité des prédictions. Par ailleurs, GeoNLPlify pourra également être évalué sur d'autres tâches de TALN (comme l'extraction d'information). Enfin, et afin de mieux comprendre l'apport de l'information spatiale, GeoNLPlify pourra être utilisé sur des corpus dont la dimension spatiale ne semble pas être fondamentale.

Remerciements : Cette étude a été partiellement financée par la subvention européenne 874850 MOOD. Le contenu de cette publication relève de la seule responsabilité des auteurs et ne reflète pas nécessairement les vues de la Commission européenne.

Références

Arsevska, E., S. Valentin, et J. Rabatel (2018). Web monitoring of emerging animal infectious diseases integrated in the French Animal Health Epidemic Intelligence System. *PLOS ONE 13*(8), 25.

Bayer, M., M.-A. Kaufhold, et C. Reuter (2022). A Survey on Data Augmentation for Text Classification. *ACM Computing Surveys 1*, 3544558.

Buntain, C., R. McCreadie, et I. Soboroff (2020). Incident Streams 2020 : TRECIS in the Time of COVID-19. In *18th International Conference on Information Systems for Crisis Response and Management*, Volume 18, pp. 621–639. ISCRAM.

9. https ://github.com/remydecoupes/GeoNLPlify

Devlin, J., M.-W. Chang, K. Lee, et K. Toutanova (2019). BERT : Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the NACL*, Volume 1, pp. 4171–4186. Association for Computational Linguistics.

Imran, M., P. Mitra, et C. Castillo (2016). Twitter as a Lifeline : Human-annotated Twitter Corpora for NLP of Crisis-related Messages. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC)*, Volume 1, pp. 1638–1643. European Language Resources Association (ELRA). arXiv :1605.05894 [cs].

Kobayashi, S. (2018). Contextual Augmentation : Data Augmentation by Words with Paradigmatic Relations. arXiv :1805.06201 [cs].

Li, Z.-z., D.-w. Feng, D.-s. Li, et X.-c. Lu (2020). Learning to select pseudo labels : a semi-supervised method for named entity recognition. *Frontiers of Information Technology & Electronic Engineering 21*(6), 903–916.

Liu, Y., M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, et V. Stoyanov (2019). RoBERTa : A Robustly Optimized BERT Pretraining Approach. Number : arXiv :1907.11692 arXiv :1907.11692 [cs].

Longpre, S., Y. Wang, et C. DuBois (2020). How Effective is Task-Agnostic Data Augmentation for Pretrained Transformers ? arXiv :2010.01764 [cs, stat].

Ma, E. (2019). nlpaug. original-date : 2019-03-21T03 :00 :17Z.

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

Morris, J. X., E. Lifland, J. Y. Yoo, J. Grigsby, D. Jin, et Y. Qi (2020). TextAttack : A Framework for Adversarial Attacks, Data Augmentation, and Adversarial Training in NLP. arXiv :2005.05909 [cs].

Ribeiro, M. T., S. Singh, et C. Guestrin (2016). "Why Should I Trust You ?" : Explaining the Predictions of Any Classifier. Number : arXiv :1602.04938 arXiv :1602.04938 [cs, stat].

Sennrich, R., B. Haddow, et A. Birch (2016). Improving Neural Machine Translation Models with Monolingual Data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, Berlin, Germany, pp. 86–96. Association for Computational Linguistics.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, et I. Polosukhin (2017). Attention Is All You Need. Number : arXiv :1706.03762 arXiv :1706.03762 [cs].

Wei, J. et K. Zou (2019). EDA : Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks. arXiv :1901.11196 [cs].

Summary

This paper proposes to use the spatial information to augment the training corpus of BERT-based text classification models on crisis related corpora. After having shown the importance of this kind of information thanks to a neural network explicability method, we propose GeoNLPlify, a set of three data augmentation techniques based on spatial information.