Méta-apprentissage : classification de messages en catégories émotionnelles inconnues en entraînement

Gaël Guibon, Matthieu Labeau, Hélène Flamein, Luce Lefeuvre, Chloé Clavel

▶ **To cite this version:**

Gaël Guibon, Matthieu Labeau, Hélène Flamein, Luce Lefeuvre, Chloé Clavel. Méta-apprentissage : classification de messages en catégories émotionnelles inconnues en entraînement. Traitement Automatique des Langues Naturelles, 2021, Lille, France. pp.199-208. hal-03265871

HAL Id: hal-03265871
https://hal.science/hal-03265871

Submitted on 23 Jun 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Actes de la 28e Conférence sur le Traitement Automatique des Langues Naturelles
Lille, France, 28 juin au 2 juillet 2021

Méta-apprentissage : classification de messages en catégories émotionnelles inconnues en entraînement

Gaël Guibon[1, 2] Matthieu Labeau[1]
Hélène Flamein[2] Luce Lefeuvre[2] Chloé Clavel[1]
(1) LTCI, Télécom-Paris, Institut Polytechnique de Paris
(2) Direction Innovation & Recherche SNCF
prénom.nom@telecom-paris.fr, prénom.nom@sncf.fr

RÉSUMÉ

Dans cet article nous reproduisons un scénario d'apprentissage selon lequel les données cibles ne sont pas accessibles et seules des données connexes le sont. Nous utilisons une approche par méta-apprentissage afin de déterminer si les méta-informations apprises à partir de messages issus de médias sociaux, finement annotés en émotions, peuvent produire de bonnes performances une fois utilisées sur des messages issus de conversations, étiquetés en émotions avec une granularité différente. Nous mettons à profit l'apprentissage sur quelques exemples (*few-shot learning*) pour la mise en place de ce scénario. Cette approche se montre efficace pour capturer les méta-informations d'un jeu d'étiquettes émotionnelles pour prédire des étiquettes jusqu'alors inconnues au modèle. Bien que le fait de varier le type de données engendre une baisse de performance, notre approche par méta-apprentissage atteint des résultats décents comparés au référentiel d'apprentissage supervisé.

ABSTRACT

Meta-learning : Classifying Messages into Unseen Emotional Categories

In this paper, we place ourselves in a classification scenario in which the target data set classes and data type are not accessible during training. We use a meta-learning approach to determine whether or not meta-trained information from common social network data with fine-grained emotion labels can achieve competitive performance on conversation utterances labeled with different, higher level, emotions. We leverage few-shot learning to concur with the classification scenario. This approach proves to be effective for capturing meta-information from a source emotional tag set to predict previously unseen emotional tags. Even though shifting the data type triggers an expected performance drop, our meta-learning approach achieves decent results when compared to the fully supervised one.

MOTS-CLÉS : classification en émotions, méta-apprentissage, apprentissage sur peu d'exemples.

KEYWORDS: emotion classification, meta-learning, few shot learning, natural language processing.

1 Introduction

Le Traitement Automatique du Langage (TAL) fait souvent face à la nécessité de devoir entraîner un modèle de classification pour une tâche sans pour autant en posséder les annotations dédiées. C'est davantage le cas pour les entreprises dont les données spécialisées, souvent privées ou confidentielles, sont synonymes d'un processus d'annotation fastidieux et coûteux. Les données annotées en émotion entrent bien souvent dans ce cas de figure puisqu'elles sont majoritairement produites dans des

Volume 1 : conférence principale, pages 199–208.
Cette œuvre est mise à disposition sous licence Attribution 4.0 International.

conversations privées et restent délicates à annoter en raison de leur subjectivité. Nos travaux se concentrent sur la tâche de classification d'émotions de textes courts et informels pour lesquels nous faisons l'hypothèse qu'un méta-apprentissage puisse servir à la classification de textes qui divergent en structure langagière et en jeu d'étiquettes.

Classifier du texte en émotions fait l'objet de nombreux travaux allant de la classification en polarités (Strapparava & Mihalcea, 2007; Thelwall *et al.*, 2012; Yadollahi *et al.*, 2017) à l'usage de représentations émotionnelles plus précises et complexes (Alm *et al.*, 2005; Bollen *et al.*, 2009; Yu *et al.*, 2015; Zhang *et al.*, 2018a; Zhu *et al.*, 2019; Zhong *et al.*, 2019; Park *et al.*, 2019). Ces approches requièrent habituellement l'usage du maximum de données possible pour entraîner le modèle de classification. Toutefois, l'obtention de jeux de données conséquents n'est pas toujours possible pour une tâche dédiée. Le recours à des stratégies telles que l'apprentissage à partir de quelques exemples (*Few Shot Learning* – FSL) (Lake, 2015; Vinyals *et al.*, 2016; Ravi & Larochelle, 2016) ou le méta-apprentissage (Schmidhuber, 1987) semble alors nécessaire. Le méta-apprentissage consiste à "apprendre à apprendre", notamment en extrayant des méta-informations permettant une application plus efficace à de nouvelles tâches.

Ces deux approches ont émergé de la vision artificielle et y ont fait l'objet de différentes stratégies d'optimisation telles que l'apprentissage épisodique (Ravi & Larochelle, 2016), le méta-apprentissage indépendant du modèle (Finn *et al.*, 2017), ou encore l'apprentissage de métrique. Cette méthode a notamment donné lieu à plusieurs modèles, comme les réseaux siamois (Koch *et al.*, 2015), les réseaux concordants (Vinyals *et al.*, 2016), et les réseaux prototypiques (Snell *et al.*, 2017). Parmi ces approches, certaines ont été adaptées à des tâches du TAL (Bao *et al.*, 2020; Gao *et al.*, 2019b), et plus particulièrement à la classification de textes. Le méta-apprentissage à l'aide du FSL a notamment été utilisé pour entraîner un modèle de classification en sentiments en variant les 23 sujets du jeu de données d'Amazon (ARSC) (Yu *et al.*, 2018; Geng *et al.*, 2019; Bao *et al.*, 2020; Bansal *et al.*, 2020). Afin d'appliquer le méta-apprentissage à la classification en émotions, de multiples approches ont été récemment abordées. Nous pouvons notamment citer une approche par apprentissage de distribution (Zhang *et al.*, 2018b) à l'aide de la décomposition de plongements de phrases, associée aux K-voisins (KNN) les plus proches (Zhao & Ma, 2019), ainsi qu'une étude de l'ambiguïté des émotions par méta-apprentissage à l'aide de BiLSTM avec attention (Fujioka *et al.*, 2019).

Dans cet article nous nous plaçons dans une situation où nous n'avons pas accès aux données cibles et par conséquent cherchons à méta-apprendre sur des données que nous savons plus ou moins proches. Nous combinons méta-apprentissage et FSL pour prédire les émotions dans des énoncés informels issus de conversations de médias sociaux, et cherchons à vérifier que cela peut nous permettre d'obtenir un modèle performant, bien qu'appris sur des données différentes. De récents travaux ont montré que le méta-apprentissage peut s'appliquer en variant les thèmes abordés des commentaires Amazon (Bao *et al.*, 2020) ou les différentes relations entre entités du corpus dédié Few-Rel (Han *et al.*, 2018; Gao *et al.*, 2019a). Dans nos travaux, nous exploitons le méta-apprentissage non seulement lorsque le jeu d'étiquettes émotionnelles diffère, mais également lorsque le format et la structure des données diffèrent. Nous contribuons en implémentant un méta-apprentissage distinguant les données à la fois par leurs jeux d'étiquettes et par leur format. Avec ce scénario, nous montrons que le méta-apprentissage combiné à l'apprentissage à partir de peu d'exemples (FSL) peut être efficace sans avoir observé ces deux variables pendant l'entraînement. Le code est publiquement disponible sur Github : `https://github.com/gguibon/metalearning-emotion-datasource`.

2 Données et étiquettes

Nous considérons deux jeux de données différents, en anglais, afin de nous conformer à la mise en place d'un métamodèle appris sur des données différentes de celles sur lesquelles il sera évalué. Nous vérifions ainsi la qualité des méta-informations apprises et la capacité de transfert du métamodèle.

GoEmotions (Demszky *et al.*, 2020) est le jeu de données sources servant à l'apprentissage de méta-informations à partir d'étiquettes et de formats différents. Il s'agit d'un corpus de 58 000 commentaires Reddit étiquetés avec 27 catégories d'émotions, catégories que nous séparons en trois jeux d'étiquettes (*EmoTagSets*) afin de permettre un méta-apprentissage par la suite.

DailyDialog (Li *et al.*, 2017) représente le jeu de données cibles à étiqueter à l'aide de notre métamodèle appris en amont. Ce corpus est constitué de 13 118 conversations de "tous les jours" sur différents thèmes. Pour nos travaux, nous n'utilisons que les messages individuels sans contexte conversationnel. Pour appliquer notre modèle nous utilisons uniquement les couples message/étiquette issus de l'échantillon de test officiel [1], soit un ensemble de 1 419 messages pour 6 émotions (*EmoTagSet3* ci-après). DailyDialog est également fourni avec une séparation en trois échantillons par *ratio* train/val/test qui assurent la répartition de toutes les étiquettes. Bien que ces échantillons par *ratio* ne permettent pas de méta-apprentissage, nous les utilisons pour de l'apprentissage supervisé standard à des fins de comparaison avec le méta-apprentissage, ainsi qu'à l'optimisation des hyperparamètres.

Jeux d'étiquettes. Nous considérons 3 jeux d'étiquettes émotionnelles distinctes, construites en prenant en compte leur nombre d'occurrences total pour avoir des tailles plus ou moins similaires, ainsi que leur présence ou non dans le corpus de test. Pour le méta-entraînement, nous utilisons le jeu nommé *EmoTagSet1* constitué des étiquettes suivantes : *admiration, approval, annoyance, amusement, love, confusion, realization, excitement, remorse, nervousness* et *pride*. Pour la validation, nous nommons *EmoTagSet2* le jeu d'étiquettes suivantes : *gratitude, curiosity, disapproval, optimism, disappointment, caring, desire, embarrassment, relief* et *grief*. Pour le test, aussi bien sur GoEmotions que sur DailyDialog, le jeu nommé *EmoTagSet3* est constitué de 6 émotions dites "basiques" : *joy, sadness, anger, fear, disgust* et *surprise*.

Ainsi, les 27 étiquettes sont séparées en 3 jeux distincts. Mis ensemble, les jeux d'entraînement et de validation représentent 21 émotions exclusives à GoEmotions, tandis que les 6 émotions du jeu de test sont présentes aussi bien dans GoEmotions que dans DailyDialog et représentent des émotions "basiques" ayant une granularité que l'on peut considérer comme moins fine. Ce dernier jeu d'étiquettes rend également la comparaison des résultats possible.

3 Méthode et protocole expérimental

GoEmotions (Demszky *et al.*, 2020) est utilisé pour l'entraînement tandis que DailyDialog (Li *et al.*, 2017) l'est pour l'évaluation, jouant ainsi le rôle de données privées non étiquetées. L'objectif est de transférer les méta-informations apprises à partir de commentaires Reddit vers des messages de conversations de tous les jours, les deux sources divergeant ainsi en structure et en vocabulaire.

Méta-apprentissage. Dans un premier temps, nous mettons en place un méta-apprentissage d'émo-

1. L'étude du contexte conversationnel dépasse le cadre de cet article. Nous nous concentrons uniquement sur les messages, les différences de structure et de jeu d'étiquettes émotionnelles pour du méta-apprentissage.

tions à partir des échantillons d'entraînement et de validation de GoEmotions. Cela nous permet d'apprendre des méta-informations que nous évaluons ensuite à l'aide de l'échantillon de test de DailyDialog, suivant le principe présenté en Figure 1. Notre but est de méta-entraîner un modèle de classification à partir de quelques exemples : plus précisément, en utilisant du FSL avec seulement 5 exemples par classe émotionnelle issue du jeu d'entraînement de GoEmotions. Nous distinguons alors 3 jeux d'étiquettes distinctes : un pour l'entraînement (*EmoTagSet1*), un pour la validation (*EmoTagSet2*) et un pour le test (*EmoTagSet3*). Les classes utilisées en test correspondent aux 6 émotions présentes dans DailyDialog afin de les exclure de l'apprentissage et de pouvoir comparer les résultats.
Pour concevoir notre méta-modèle, nous utilisons les réseaux prototypiques (Snell *et al.*, 2017) avec un apprentissage épisodique (Ravi & Larochelle, 2016) afin de combiner méta-apprentissage et FSL.

Concrètement, un épisode est défini par trois paramètres : le nombre de classes N_c (*ways*), le nombre d'exemples d'entraînement N_s (*shots*) pour chaque classe et le nombre d'éléments à étiqueter N_q (*queries*). Les éléments de l'épisode sont projetés dans un espace vectoriel, dans lequel sont calculés les prototypes de classes, à l'aide d'un encodeur f_ϕ. Lors de chaque épisode, nous appliquons l'apprentissage de métrique aux quelques exemples visibles S_k (*shots*) d'une classe k (*way*) pour lui attribuer un prototype $\mathbf{c}_k$ qui correspond à la moyenne des exemples $\mathbf{x}_i \in S_k$ une fois encodés par f_ϕ. Un prototype de classe est donc égal à : $\mathbf{c}_k \leftarrow \frac{1}{N_C} \sum_{(\mathbf{x}_i, y_i) \in S_k} f_\phi(\mathbf{x}_i)$ où N_C représente le nombre de classes. La fonction d'encodage f_ϕ est ensuite également appliquée aux éléments à étiqueter, soit le jeu de requêtes Q_k. Nous minimisons ensuite la distance euclidienne d entre les vecteurs des prototypes de chaque classe et les vecteurs issus de l'encodage des éléments à étiqueter $d\left(f_\phi(\mathbf{x}), \mathbf{c}_k\right)$. L'encodeur est mis à jour à l'aide du calcul de l'erreur suivant : $\frac{1}{N_C N_Q}\left[d\left(f_\phi(\mathbf{x}), \mathbf{c}_k\right) + \log \sum_{k'} \exp\left(-d\left(f_\phi(\mathbf{x}), \mathbf{c}_{k'}\right)\right)\right]$ où N_Q représente le nombre d'éléments à étiqueter (communément nommé *Query Set*). Un épisode se résume donc à créer un prototype pour chaque classe à partir de peu d'exemples avant d'assigner une classe à un élément à étiqueter en fonction du prototype de classe qui lui est le plus proche. Les réseaux prototypiques créent ainsi, par le biais de l'encodeur, des vecteurs représentant chaque classe à partir de quelques exemples du jeu de support S (*Support Set*), correspondant au jeu d'exemples aléatoires dans lequel nous obtenons N_s exemples pour chaque classe N_c. Dans le cas du méta-apprentissage, ces vecteurs, ou prototypes, sont ici utilisés pour prédire des classes jamais vues dans le jeu d'entraînement. L'encodeur est ainsi optimisé pour chercher à obtenir une représentation de méta-informations inhérentes à chaque classe.

Nous varions les encodeurs en considérant alternativement : la moyenne des plongements lexicaux des éléments de la classe (AVG), un encodeur convolutionnel (CNN) (Kim, 2014) ou un encodeur *Transformer* (Vaswani *et al.*, 2017) (Transfo). La Figure 1 montre une vision globale de notre stratégie de méta-apprentissage, de l'apprentissage à l'évaluation. Nous définissons la composition d'un épisode par $N_c = 6$, $N_s = 5$ et $N_q = 30$, signifiant alors une tâche d'apprentissage en 5 exemples aléatoires, 6 classes et 30 cibles aléatoires (*5-shot 6-way 30-query*) dans laquelle N_c est contraint par le nombre de classes en test, le modèle étant *in fine* testé sur les 6 émotions du jeu de test de DailyDialog.

La composition des épisodes pour l'entraînement et pour la validation est identique. Pour le test, en revanche, nous utilisons 1 000 épisodes aléatoires pour lesquels le jeu d'exemples cibles (*query set*) est choisi aléatoirement à partir du jeu de test tout en suivant le principe des 5 exemples cibles par classe (c'est-à-dire 5 exemples à étiqueter pour chacune des 6 émotions). Afin d'assurer l'évaluation du méta-apprentissage, ces 6 classes sont disponibles uniquement lors du test.

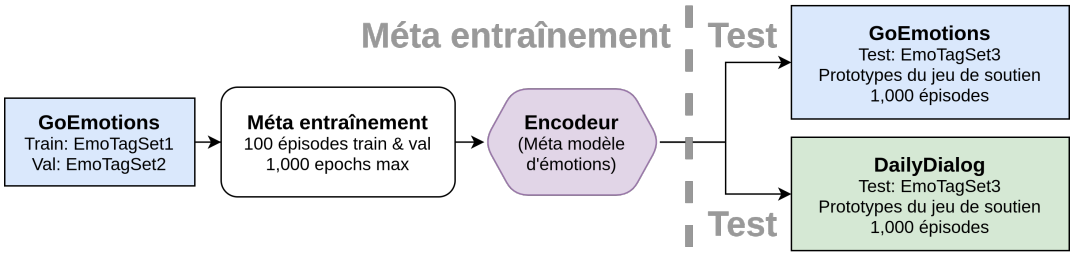

FIGURE 1 – Vue globale de la stratégie de méta-apprentissage. Lors du test sur DailyDialog, seuls les messages du jeu de test officiel sont pris en compte. EmoTagSet1 ∩ EmoTagSet2 ∩ EmoTagSet3 = ∅.

Notre protocole expérimental est le suivant : chaque jeu de données est pré-traité à l'aide du `TweetTokenizer` de NLTK [2] et de la normalisation en minuscules puis, pour les représentations textuelles de départ, nous utilisons les plongements lexicaux d'un modèle pré-entraîné de *fastText* (Joulin *et al.*, 2017), suivant ainsi des travaux antérieurs (Bao *et al.*, 2020). Alternativement, nous utilisons un modèle BERT pré-entraîné (Devlin *et al.*, 2019), bien que ce dernier n'améliore guère les résultats, confortant l'hypothèse de l'impact réduit sur des données de messages non séparés en phrases et qui possèdent une variation excessive du nombre de *tokens* (Bao *et al.*, 2020).

Apprentissage supervisé pour comparaison future. Dans un premier temps nous appliquons un apprentissage supervisé standard en utilisant uniquement les jeux de données officiels par *ratio* de DailyDialog (train/val/test). L'apprentissage supervisé sert donc de référence, démontrant les scores atteignables avec une approche classique et permettant ensuite une comparaison avec le méta-apprentissage. Dans cette approche supervisée, l'encodeur et le modèle de classification ne sont pas distincts puisqu'il s'agit simplement d'ajouter une couche cachée affine suivie d'un *softmax*, puis de calculer la log-vraisemblance négative pour obtenir une entropie croisée sur les différentes prédictions.

Hyper-paramètres. Pour rappel, nous répliquons ici un scénario dans lequel nous entraînons un modèle de classification sans avoir accès aux données cibles. Cependant, à des fins de comparaison, nous utilisons les hyper-paramètres obtenus à l'aide d'une recherche ciblée par quadrillage effectuée uniquement lors de l'apprentissage supervisé. Cela réduit la dépendance des résultats à d'éventuels paramètres spécifiques au méta-apprentissage et implique une évaluation plus fiable. Les hyper-paramètres généraux sont un pas d'apprentissage de $1e-3$, des vecteurs en entrée de taille 300, un abandon (*dropout*) de 10% sans coupe de gradient et un arrêt de l'apprentissage au bout de 20 *epochs* (*i.e.* un ensemble de 100 épisodes aléatoires). Les CNN suivent la configuration de Kim (Kim, 2014) avec des filtres ayant trois tailles différentes, de 3 à 5. Contrairement à Kim, nous utilisons ici 5 000 filtres au lieu de 50. Pour l'encodeur en *Transformer* le pas d'apprentissage est de $1e-4$, le *dropout* est augmenté à 20% et celui de l'encodeur positionnel est conservé à 10%.

Métriques d'évaluation. Nous évaluons les performances des modèles en nous inspirant de travaux précédents du FSL qui utilisent l'exactitude (Snell *et al.*, 2017; Sung *et al.*, 2018; Bao *et al.*, 2020). Toutefois, nous allons plus loin en considérant également une F1-mesure pondérée et le coefficient de corrélation de Matthews (MCC) (Cramir, 1946; Baldi *et al.*, 2000) comme suggéré par de récentes études en biologie (Chicco & Jurman, 2020), mais dans une version adaptée à une classification multi-classes (Gorodkin, 2004) qui convient davantage à notre tâche. Chaque résultat affiché représente la moyenne de l'ensemble des épisodes de test.

2. https://www.nltk.org/

4 Résultats

Apprentissage supervisé (hyperparamètres)
DailyDialog (6 classes)

Enc.	Clf.	Acc	F1	MCC
AVG	MLP	49,73	42,06	42,32
CNN	MLP	**62,57**	**54,89**	**59,12**
Transfo	MLP	55,35	48,52	49,24

Méta-apprentissage
GoEmotions : 6 way 5 shot 30 query

Enc.	Clf.	Acc ±	F1 ±	MCC ±
AVG	Proto	39,00 04, 8	38,35 04, 9	27,14 05, 8
CNN	Proto	42,83 04, 9	42,11 05, 0	31,76 05, 9
Dist.	Ridge	43,67 09, 8	42,71 09, 5	33,20 12, 0
Transfo	Proto	**95,89** 04, 3	**95,27** 05, 3	**95,31** 04, 8
Évaluation du métamodèle sur le jeu de test de DailyDialog : 1 000 épisodes				
AVG	Proto	19,60 03, 6	19,38 03, 7	03,40 04, 1
CNN	Proto	19,37 03, 6	18,77 03, 7	03,30 04, 3
Dist.	Ridge	26,05 08, 1	24,78 07, 9	11,58 10, 0
Transfo	Proto	**46,29** 25, 8	**39,02** 29, 6	**40,67** 30, 7

TABLE 1 – À gauche : apprentissage supervisé sur les messages de DailyDialog ; À droite : le méta-apprentissage entraîné en séparant les classes de GoEmotions, puis appliqué sur DailyDialog (test) ; évalués en exactitude (Acc), F1-mesure et coefficient de corrélation de Matthews (MCC). ± montre la variance au fil des épisodes.

Résultats de l'apprentissage supervisé. Les résultats présentés en partie gauche du tableau 1 proviennent de l'utilisation des jeux de données officiels de DailyDialog. Comme expliqué en Section 3, nous cherchons les meilleurs hyperparamètres pour chaque encodeur et chaque modèle de classification lors de cette phase d'apprentissage supervisé. Ceci s'avère nécessaire et sensible en particulier concernant les *transformers* (Vaswani *et al.*, 2017) qui requièrent un ajustement précis pour converger, d'autant plus quand, comme c'est le cas ici, le jeu de données est de taille relativement petite. La taille du jeu de données semble précisément être la raison pour laquelle le modèle de classification par *transformers* donne des résultats inférieurs à ceux du CNN dans cette approche purement supervisée.

Résultats du méta-apprentissage. La seconde section du tableau 1 présente deux résultats principaux : ceux de la phase de méta-apprentissage sur GoEmotions utilisant les jeux de données par ensemble d'étiquettes (11 émotions en entraînement, 10 autres en validation et 6 différentes en test) et l'évaluation de ces modèles sur le jeu de test officiel de DailyDialog (6 émotions). Comme on peut s'y attendre, le méta-apprentissage donne de moins bons résultats que l'apprentissage supervisé. Cela s'explique par l'apprentissage de méta-informations sur des sources variant en jeu d'étiquettes, en longueurs de phrases et ayant un contexte conversationnel différent. Les résultats obtenus montrent qu'une structure linguistique similaire entre le jeu de données cibles et celui d'entraînement facilite l'apprentissage de méta-informations, entraînant alors de meilleures performances. En effet, les résultats du méta-apprentissage obtenus sur GoEmotions sont meilleurs que ceux obtenus sur DailyDialog, bien qu'il s'agisse du même modèle. D'autre part, contrairement aux résultats de l'approche supervisée, l'utilisation d'un *Transformer* en tant qu'encodeur, ici associé aux réseaux prototypiques pour permettre un méta-entraînement, surpasse les autres encodeurs de manière significative. Cette approche surpasse également la baseline récente associant signatures de distribution (Dist.) sous forme d'attention à un Ridge Regressor (Ridge) (Bao *et al.*, 2020). Nous pensons que les piètres résultats obtenus en utilisant les CNN (Kim, 2014) en tant qu'encodeur démontrent le besoin d'attention dans le processus d'apprentissage de méta-informations pertinentes, d'autant plus avec une quantité réduite de données, ce qui confirmerait les précédents travaux faisant une observation similaire (Sun *et al.*, 2019).

5 Discussions

Méta-modèles et fonctionnement sur des émotions inconnues en entraînement. Les réseaux prototypiques utilisent le jeu de support (*support set*) pour calculer un prototype pour chaque classe (*way*), créant ainsi de nouveaux prototypes lors de chaque épisode. Cela signifie que l'encodeur entraîné ne dépend pas des classes prédites mais des informations regroupées déterminant la position des éléments dans l'espace vectoriel. La proximité permettant d'assigner un élément cible (*query*) à un prototype de classe est relative et, de ce fait, encoder un élément "loin" des prototypes dans l'espace vectoriel n'aura pas nécessairement d'impact sur la qualité de la prédiction.

Nature et ambiguïté des étiquettes émotionnelles. Il existe des liens non systématiques d'hyperonymie entre les émotions de base (*EmoTagSet3*) et les 21 émotions à granularité fine exclusives à GoEmotions et utilisées pour l'entraînement et la validation. Ces relations, fournies dans GoEmotions, peuvent être contestées. Ainsi, la joie est définie comme ayant notamment pour sous-catégories l'amusement et l'approbation ; tandis que la surprise inclut la curiosité. En outre, avec 95,27 % en F1-mesure, le méta-apprentissage fonctionne très bien sur GoEmotions (tableau 1) et semble ainsi prendre en compte l'ambiguïté et la différence de granularité des étiquettes.

Le méta-apprentissage à partir de différentes sources. Nous avons effectué un perfectionnement (*fine-tuning*) des modèles appris sur GoEmotions en utilisant le jeu de test de ce corpus (6 émotions), avant de l'appliquer sur DailyDialog. Ce perfectionnement était constitué d'une itération de 10 épisodes supplémentaires, contrairement aux 1 000 itérations maximales sur 100 épisodes utilisées lors de l'entraînement, et avait pour objectif de légèrement adapter l'encodeur au jeu d'étiquettes cibles en tirant parti des méta-informations apprises lors de l'entraînement. Le jeu d'étiquettes en test étant constitué des mêmes 6 émotions, ce perfectionnement nous permet principalement de vérifier si la difficulté de la tâche est due à la variété du jeu d'étiquettes ou à la variété des sources de données, qu'il s'agisse de GoEmotions (test) ou DailyDialog (test). Cette procédure a fourni des résultats de qualité moindre, nous conduisant à l'hypothèse selon laquelle les différences entre les structures langagières (messages de réseaux sociaux et messages informels journaliers) sont la principale source d'erreurs lors de l'utilisation de notre méta-apprentissage.

6 Conclusion

Nous avons mis en place un scénario de classification dans lequel nous ne possédons qu'un seul type de données d'entraînement, sans garantie de la similarité des données de test aussi bien au niveau du format que du jeu d'étiquettes. Nous avons utilisé des messages issus de médias sociaux étiquetés en émotions fines pour l'obtention de méta-informations à l'aide de la combinaison de méta-entraînement et d'entraînement à partir de peu d'exemples, évaluée sur des messages issus de conversations et comportant un jeu d'étiquettes différent. Pour cela nous avons mis en place des réseaux prototypiques avec pour encodeur un *Transformer*, appris par approche épisodique afin de simuler l'accès restreint aux données. Nous avons ainsi obtenu des résultats encourageants si l'on compare les performances entre le méta-modèle et le modèle de référence appris de manière supervisée. Cette approche fonctionne dans le cas de l'apprentissage de méta-informations liées aux différentes émotions mais peine à s'adapter à la variation de source de données. Par la suite, nous souhaitons vérifier dans de futurs travaux si le méta-entraînement peut être étendu à un cadre de classification utilisant le contexte conversationnel précédent un message.

Références

ALM C. O., ROTH D. & SPROAT R. (2005). Emotions from text : machine learning for text-based emotion prediction. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing - HLT '05*, p. 579–586, Vancouver, British Columbia, Canada : Association for Computational Linguistics. DOI : 10.3115/1220575.1220648.

BALDI P., BRUNAK S., CHAUVIN Y. & NIELSEN H. (2000). Assessing the accuracy of prediction algorithms for classification : An overview. *Bioinformatics*, **16**(5), 412–424. DOI : 10.1093/bioinformatics/16.5.412.

BANSAL T., JHA R., MUNKHDALAI T. & MCCALLUM A. (2020). Self-supervised meta-learning for few-shot natural language classification tasks. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, p. 522–534, Online : Association for Computational Linguistics.

BAO Y., WU M., CHANG S. & BARZILAY R. (2020). Few-shot text classification with distributional signatures. In *International Conference on Learning Representations*.

BOLLEN J., PEPE A. & MAO H. (2009). Modeling public mood and emotion : Twitter sentiment and socio-economic phenomena. *arXiv :0911.1583 [cs]*. arXiv : 0911.1583.

CHICCO D. & JURMAN G. (2020). The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. *BMC genomics*, **21**(1), 6.

CRAMIR H. (1946). Mathematical methods of statistics. *Princeton U. Press, Princeton*, **500**.

DEMSZKY D., MOVSHOVITZ-ATTIAS D., KO J., COWEN A., NEMADE G. & RAVI S. (2020). GoEmotions : A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, p. 4040–4054, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.acl-main.372.

DEVLIN J., CHANG M.-W., LEE K. & TOUTANOVA K. (2019). BERT : Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long and Short Papers)*, p. 4171–4186, Minneapolis, Minnesota : Association for Computational Linguistics. DOI : 10.18653/v1/N19-1423.

FINN C., ABBEEL P. & LEVINE S. (2017). Model-agnostic meta-learning for fast adaptation of deep networks. In D. PRECUP & Y. W. TEH, Éds., *Proceedings of the 34th International Conference on Machine Learning*, volume 70 de *Proceedings of Machine Learning Research*, p. 1126–1135, International Convention Centre, Sydney, Australia : PMLR.

FUJIOKA T., BERTERO D., HOMMA T. & NAGAMATSU K. (2019). Addressing ambiguity of emotion labels through meta-learning. *CoRR*, **abs/1911.02216**.

GAO T., HAN X., LIU Z. & SUN M. (2019a). Hybrid Attention-Based Prototypical Networks for Noisy Few-Shot Relation Classification. *Proceedings of the AAAI Conference on Artificial Intelligence*, **33**, 6407–6414. DOI : 10.1609/aaai.v33i01.33016407.

GAO T., HAN X., ZHU H., LIU Z., LI P., SUN M. & ZHOU J. (2019b). FewRel 2.0 : Towards More Challenging Few-Shot Relation Classification. *arXiv :1910.07124 [cs]*. arXiv : 1910.07124.

GENG R., LI B., LI Y., ZHU X., JIAN P. & SUN J. (2019). Induction Networks for Few-Shot Text Classification. *arXiv :1902.10482 [cs]*. arXiv : 1902.10482.

GORODKIN J. (2004). Comparing two k-category assignments by a k-category correlation coefficient. *Computational biology and chemistry*, **28**(5-6), 367–374.

HAN X., ZHU H., YU P., WANG Z., YAO Y., LIU Z. & SUN M. (2018). FewRel : A Large-Scale Supervised Few-Shot Relation Classification Dataset with State-of-the-Art Evaluation. *arXiv :1810.10147 [cs, stat]*. arXiv : 1810.10147.

JOULIN A., GRAVE E., BOJANOWSKI P. & MIKOLOV T. (2017). Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics : Volume 2, Short Papers*, p. 427–431, Valencia, Spain : Association for Computational Linguistics.

KIM Y. (2014). Convolutional neural networks for sentence classification. *arXiv preprint arXiv :1408.5882*.

KOCH G., ZEMEL R. & SALAKHUTDINOV R. (2015). Siamese Neural Networks for One-shot Image Recognition. *ICML*, p.$\tilde{8}$.

LAKE B. (2015). LakeEtAl2015Science-startOfFewShot.pdf. *Sciences Mag*.

LI Y., SU H., SHEN X., LI W., CAO Z. & NIU S. (2017). Dailydialog : A manually labelled multi-turn dialogue dataset. In *Proceedings of The 8th International Joint Conference on Natural Language Processing (IJCNLP 2017)*.

PARK S., KIM J., JEON J., PARK H. & OH A. (2019). Toward dimensional emotion detection from categorical emotion annotations. *arXiv preprint arXiv :1911.02499*.

RAVI S. & LAROCHELLE H. (2016). Optimization as a model for few-shot learning.

SCHMIDHUBER J. (1987). *Evolutionary principles in self-referential learning, or on learning how to learn : the meta-meta-... hook*. Thèse de doctorat, Technische Universität München.

SNELL J., SWERSKY K. & ZEMEL R. (2017). Prototypical networks for few-shot learning. In *Advances in neural information processing systems*, p. 4077–4087.

STRAPPARAVA C. & MIHALCEA R. (2007). Semeval-2007 task 14 : Affective text. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, p. 70–74.

SUN S., SUN Q., ZHOU K. & LV T. (2019). Hierarchical Attention Prototypical Networks for Few-Shot Text Classification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, p. 476–485, Hong Kong, China : Association for Computational Linguistics. DOI : 10.18653/v1/D19-1045.

SUNG F., YANG Y., ZHANG L., XIANG T., TORR P. H. & HOSPEDALES T. M. (2018). Learning to compare : Relation network for few-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, p. 1199–1208.

THELWALL M., BUCKLEY K. & PALTOGLOU G. (2012). Sentiment strength detection for the social web. *Journal of the American Society for Information Science and Technology*, **63**(1), 163–173.

VASWANI A., SHAZEER N., PARMAR N., USZKOREIT J., JONES L., GOMEZ A. N., KAISER Ł. & POLOSUKHIN I. (2017). Attention is all you need. In *Advances in neural information processing systems*, p. 5998–6008.

VINYALS O., BLUNDELL C., LILLICRAP T., WIERSTRA D. *et al.* (2016). Matching networks for one shot learning. In *Advances in neural information processing systems*, p. 3630–3638.

YADOLLAHI A., SHAHRAKI A. G. & ZAIANE O. R. (2017). Current state of text sentiment analysis from opinion to emotion mining. *ACM Computing Surveys (CSUR)*, **50**(2), 1–33.

YU L.-C., WANG J., LAI K. R. & ZHANG X.-J. (2015). Predicting Valence-Arousal Ratings of Words Using a Weighted Graph Method. 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)*, p. 788–793, Beijing, China : Association for Computational Linguistics. DOI : 10.3115/v1/P15-2129.

YU M., GUO X., YI J., CHANG S., POTDAR S., CHENG Y., TESAURO G., WANG H. & ZHOU B. (2018). Diverse Few-Shot Text Classification with Multiple Metrics. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long Papers)*, p. 1206–1215, New Orleans, Louisiana : Association for Computational Linguistics. DOI : 10.18653/v1/N18-1109.

ZHANG Y., FU J., SHE D., ZHANG Y., WANG S. & YANG J. (2018a). Text Emotion Distribution Learning via Multi-Task Convolutional Neural Network. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, p. 4595–4601, Stockholm, Sweden : International Joint Conferences on Artificial Intelligence Organization. DOI : 10.24963/ijcai.2018/639.

ZHANG Y., FU J., SHE D., ZHANG Y., WANG S. & YANG J. (2018b). Text emotion distribution learning via multi-task convolutional neural network. In *IJCAI*, p. 4595–4601.

ZHAO Z. & MA X. (2019). Text emotion distribution learning from small sample : A meta-learning approach. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, p. 3957–3967, Hong Kong, China : Association for Computational Linguistics. DOI : 10.18653/v1/D19-1408.

ZHONG P., WANG D. & MIAO C. (2019). Knowledge-Enriched Transformer for Emotion Detection in Textual Conversations. *arXiv :1909.10681 [cs]*. arXiv : 1909.10681.

ZHU S., LI S. & ZHOU G. (2019). Adversarial Attention Modeling for Multi-dimensional Emotion Regression. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, p. 471–480, Florence, Italy : Association for Computational Linguistics. DOI : 10.18653/v1/P19-1045.