Prédiction automatique d'emojis sentimentaux

Gaël Guibon — Magalie Ochs — Patrice Bellot

Aix Marseille Université, Université de Toulon, CNRS, ENSAM, LSIS, Marseille, France

RÉSUMÉ. *Dans les messageries sociales les emojis sont parmi les principaux vecteurs d'émotions et de sentiments des individus. Aujourd'hui, les utilisateurs naviguent dans des bibliothèques contenant souvent des milliers d'emojis pour sélectionner celui correspondant à ce qu'ils souhaitent transmettre. Nos travaux visent à développer un système de recommandation automatique d'emoji permettant à l'utilisateur d'identifier un panel réduit d'emojis pertinents étant donnée sa conversation en évitant le parcours de bibliothèques conséquentes d'emojis. Cette recommandation pouvant permettre à l'utilisateur de requêter les phrases susceptibles de contenir cet emoji, et l'émotion qui y est associée. Pour ce faire, dans un premier temps, notre objectif est de développer un outil permettant de prédire automatiquement les emojis d'une phrase à partir d'un modèle de classification appris sur un corpus de messagerie sociale contenant des emojis. Plusieurs caractéristiques sont considérées pour l'apprentissage telles que le sentiment de l'utilisateur mais aussi son humeur. Dans cet article, nous décrivons l'impact de ces caractéristiques et les performances des modèles résultants.*

ABSTRACT. *Emojis are among the main carriers of emotions and sentiment in social messaging applications. Nowadays users have to scroll down libraries of thousands of emojis in order to select the one they wanted to use. Our work aims to build an emoji automatic recommendation system to avoid scrolling emoji libraries. And which will allow the user to request emojis by the current sentence based on the emotion it conveys. To do so, we first contribute by building an emoji automatic prediction in sentences based on a classification model. This classification model is learned on an informal text messages corpus based on real data containing emojis. Several features are used to train the classifier. Such as the sentiment value of the text and the user's mood. In this paper we describe the features and models impact on the emoji prediction task.*

MOTS-CLÉS : *Classification multi-étiquette, recommandation d'emoji, analyse de sentiment.*

KEYWORDS: *Multilabel classification, emoji recommendation, sentiment analysis.*

1. Introduction

Les messageries sociales actuelles sont l'un des moyens de communication les plus utilisés avec plus de la moitié (55%) des adolescents envoyant au moins un message instantané par jour sur téléphone portable (Lenhart *et al.*, 2015). Elles se trouvent sous des formes diverses et variées, qu'il s'agisse d'une messagerie sociale non instantanée tel un forum ou un outil de micro-blogging, ou bien instantanée telle une application de messages privés ou une salle de chat.

L'utilisation des emojis a crû de façon spectaculaire depuis l'introduction du clavier emojis dans l'iOS d'Apple en 2011. Si bien que la quasi-totalité des interfaces de messageries sociales possèdent désormais la fonctionnalité d'envoi d'emojis afin de permettre à l'utilisateur d'exprimer visuellement ses émotions et sentiments. A ce jour, ce sont désormais 92% des utilisateurs en ligne qui utilisent des emojis (Team, 2015). Ces derniers remplacent au fur et à mesure une bonne partie du vocabulaire propre à la communication en ligne. Cependant, afin de sélectionner l'icône correspondant au message non-verbal qu'il souhaite transmettre, l'utilisateur est souvent amené à naviguer dans des bibliothèques contenant des milliers d'emojis parmi une liste d'emojis standards et propriétaires ne cessant de s'agrandir par centaines. Les 104 nouveaux emojis d'Apple avec l'iOS 10.2[1], ou encore les nouveaux emojis d'Android 7.1 représentant des professions[2] en sont un bon exemple. Si l'on se restreint aux emojis standards, c'est-à-dire aux emojis communément admis et utilisés dans la plupart des applications, le consortium Unicode en compte actuellement 2389 dont 797 ajoutés en 2015 et 233 depuis le début de l'année 2016[3]. Ces emojis font également l'objet de refontes graphiques régulières visant à uniformiser leur interprétation et à minimiser les différentes interprétations possibles (Miller *et al.*, 2016)(Kelly et Watts, 2015).

Nos travaux de recherche visent à *recommander automatiquement* les emojis pertinents à l'utilisateur en fonction de plusieurs paramètres. Dans l'industrie, cette recommandation est actuellement limitée à une suggestion d'emojis en fonction du mot actuel, tel iMessage sur iOS 10. L'une des stratégies possibles pour la recommandation est la prédiction d'étiquettes (De Oliveira *et al.*, 2013), mais ceci ne constitue que la première étape d'un système de recommandation complet (Avazpour *et al.*, 2014). Un système de recommandation complet ne se limite pas à recommander ce qu'il prédit par une classification. Dans cet article nous nous focalisons d'abord sur la prédiction d'emojis au niveau phrastique en abordant ce problème comme une tâche de classification multi-étiquette (dite *multi-label*), chaque classe correspondant à un emoji, chaque phrase pouvant appartenir à plusieurs classes en même temps. Reproduisant ainsi le fait qu'une phrase puisse contenir plusieurs emojis. Nous nous focalisons sur les émotions et sentiments, c'est pourquoi nous nous basons sur des approches d'analyse de sentiment lors de l'apprentissage.

1. http://emojipedia.org/apple/ios-10.2/new/
2. http://blog.emojipedia.org/android-7-1-emoji-changelog/
3. http://unicode.org/emoji/charts/full-emoji-list.html

Le papier est organisé comme suit. Nous faisons d'abord un résumé de l'état actuel des travaux de recommandation d'emojis dans les applications de messagerie sociale (Section 2). Dans la section 3 nous décrivons notre corpus de données de messagerie sociale et ses caractéristiques. En section 4 les différents modèles de classification sont appris et évalués sur le corpus. Nous concluons enfin notre approche.

Notre contribution peut se résumer aux principaux points suivants : l'apprentissage de modèles de classification pour la prédiction automatique d'emojis par approche dynamique au fil des messages, l'exploitation de données réelles de conversations privées pour l'apprentissage et le test des modèles avec l'identification de leurs caractéristiques déterminantes, et vérification de l'efficacité de la prédiction des emojis sentimentaux sur un corpus étendu.

2. Etat de l'art

Les émotions et les expressions faciales sont représentées par des caractères (*émoticônes* : :-) , :P)), ou des images (*emojis* : 😘). Les emojis permettent également de représenter des idées, puisqu'ils constituent des images et ne sont donc pas limités à un nombre de caractères contrairement aux émoticônes. Les emojis étant des images, ils ne peuvent pas être considérés comme du texte par les méthodes de Traitement Automatique du Langage (TAL) ou de Recherche d'Information (RI) classiques. Les emojis sont des signes graphiques, et en ce sens ils possèdent une signification (Peirce, 1902) qui est bien souvent dépendante du contexte d'apparition. Les emojis tendent à remplacer petit à petit les émoticônes dans les conversations sociales (Pavalanathan et Eisenstein, 2015). Ils sont également plus nombreux et plus complexes, c'est pourquoi dans nos travaux nous nous focalisons sur les emojis.

2.1. *Recommandation d'emojis*

Les emojis sont utilisés dans 70% des cas pour faciliter la compréhension d'un message (Kelly et Watts, 2015). Une bonne recommandation des emojis est donc essentielle dans un contexte de messagerie sociale, qu'elle soit publique ou privée, pour améliorer la qualité et la précision du dialogue. La tâche de classification constitue la première étape d'un système de recommandation. La classification de texte par emojis est ainsi l'objet de travaux récents.

Une classification efficace est nécessaire pour prédire les emojis et mieux les recommander. Cette classification peut prendre plusieurs formes. (Eisner *et al.*, 2016) ont par exemple effectué des plongements lexicaux (*word embeddings*) d'emojis à partir de leur description textuelle donnée par Unicode. Cette description prend la forme suivante : 😘, *"smiling face with heart-eyes"*. Cette description des emojis permettrait de les définir sans avoir avoir à prendre en compte le contexte d'apparition de l'emoji. Une classification multi-classe de description d'emojis a été ainsi faite pour évaluer

leur modèle, attribuant le bon emoji à la bonne description avec un taux d'exactitude de 85%. (Xie *et al.*, 2016) ont de leur côté également abordé la recommandation d'emoji comme un problème de classification par réseaux de neurones appris sur des données réelles issues de la plateforme sociale Weibo[4]. Cette classification a été réalisée dans un contexte conversationnel, pour une performance de 65% de précision sur les 3 emojis les plus utilisés dans leur corpus. Toute la classification se limitait ici aux 10 emojis les plus utilisés.

On peut faire un parallèle entre la classification de texte par émotions (Chaffar et Inkpen, 2011) et par humeurs (Mishne *et al.*, 2005) avec la classification de texte par emojis. En effet, prédire automatiquement une émotion ou une humeur permettrait d'identifier ensuite automatiquement un sous-ensemble d'emojis correspondants. Il serait possible de se restreindre aux 7 classes d'emotions inspirées des notions des émotions basiques d'Ekman (Ekman, 1993) que sont la colère, le dégoût, la peur, le bonheur, la tristesse et la surprise. C'est ce qu'ont fait Alm *et al.* en classant automatiquement des contes de fée (Alm *et al.*, 2005). Plus récemment, (Li et Xu, 2014) ont ainsi classifié les messages de blogs par emotions à l'aide d'une extraction de la cause de l'émotion.

Nos travaux de recherche se distinguent des travaux décrits ci-dessus. Nous nous focalisons sur une approche dynamique de prédiction au fil des messages, et considérons l'utilisation consécutive de plusieurs emojis, en plus d'aller au delà des 10 emojis les plus utilisés. Nous proposons pour cela une approche par classification multi-étiquette (*multilabel*).

2.2. *Classification multi-étiquette*

La classification multi-étiquette est une généralisation du problème de classification classique. La classification est souvent représentée par une séparation des éléments en deux classes après avoir appris à partir de données représentant qu'un seul label (classification binaire) ou plusieurs classes (classification multi-classe). La classification multi-étiquette consiste à attribuer à chaque élément une ou plusieurs étiquettes (*labels*), soit un jeu d'étiquettes pour un seul élément. Contrairement à la classification multi-classe, chaque classe n'est ici pas exclusive. La classification multi-étiquette est notamment née du besoin de ranger des objets par catégories. Récemment, elle a par exemple été utilisée pour attribuer des domaines à des documents (Rubin *et al.*, 2012), ou encore pour classer des œuvres musicales par émotions (Trohidis *et al.*, 2008).

Il y a deux principales approches de classification multi-étiquette : l'approche par transformation et celle par adaptation[5] Ces approches sont rappelées dans les récentes revues d'état de l'art de classification multi-étiquette (Tsoumakas et Katakis, 2006) (Zhang et Zhou, 2014).

4. `http://www.weibo.com/`

5. Dans certains cas, une troisième approche d'ensemble de classifieurs par transformation ou adaptation est également évoquée (Nair-Benrekia, 2015).

Multi-étiquetage par transformation. La classification multi-étiquette par transformation consiste à transformer le problème de classification multi-étiquette en un problème de classification binaire. Chaque classe fait ainsi l'objet d'un classifieur binaire attitré qui prédira si oui ou non l'individu fait partie de cette classe. Le processus est répété autant de fois qu'il y a de classes. Le résultat de la classification multi-étiquette devient alors l'addition des résultats de classification binaire de chaque classe (Lauser et Hotho, 2003).

Multi-étiquetage par adaptation. Cette approche consiste à adapter les algorithmes de classification à la tâche de classification multi-étiquette. Contrairement à l'approche précédente, un seul classifieur est appris. Elle fut par exemple utilisée pour une classification multi tâche et multi étiquette de sentiments et domaines (Huang *et al.*, 2013).

Dans notre contexte, étant donné le nombre de classes possibles (1070 au total, correspondant aux 1070 emojis de notre corpus), pour des raisons de temps de calcul et de performance, nous adoptons une approche par adaptation. Dans cette perspective, nous utilisons les *ML-RandomForest* (Breiman, 2001), les forêts aléatoires d'arbres de décision adaptées à un apprentissage multi-étiquette. Plusieurs arbres de décision sont ainsi appris aléatoirement avant d'être moyennés. Les *ML-RandomForest* sont un algorithme populaire de classification qui a pour avantage de mesurer un score d'importance des caractéristiques utilisées en plus de posséder une forte capacité de généralisation (Strobl *et al.*, 2008), évitant de ce fait le sur-apprentissage.

Dans la section suivante, nous décrivons le corpus et l'ensemble des caractéristiques disponibles. Tous deux servant par la suite à l'apprentissage de classifieurs multi-étiquettes pour prédire les emojis dans les phrases.

3. Corpus de messagerie sociale privée

Avant d'aborder la prédiction d'emojis à l'aide de classification supervisée, il convient de définir le corpus que nous utilisons pour construire et appliquer ces modèles de classification. Ces corpus sont constitués de messages informels privés de langue anglaise dont certaines caractéristiques sont illustrées dans le tableau 2. Ils proviennent initialement d'un ensemble de 1 300 000 messages confidentiels dont nous avons extrait uniquement les messages contenant des emojis pour permettre au classifieur supervisé d'apprendre les corrélations entre les emojis et les caractéristiques des phrases. Ainsi pour le premier corpus ("corpus étendu" dans le Tableau 2) nous ne récupérons que les phrases qui contiennent des emojis, et pour le second uniquement les phrases qui contiennent des emojis sentimentaux ("corpus dédié" dans le Tableau 2). Par le terme "emojis sentimentaux" nous désignons les emojis représentant des sentiments (amour, joie, tristesse, etc.). On les distingue des emojis objets tels qu'une voiture, un drapeau ou un café par exemple. La détection des phrases est effectuée

à l'aide du modèle anglais d'OpenNLP[6] (Baldridge, 2005). Le tableau 1 montre un exemple d'une phrase du corpus et de ses emojis associés.

😪😓	I heard about the news, it actually is quite depressing

Tableau 1 : Exemple factice d'une phrase représentée par la paire emojis|texte

Dans le corpus, nous avons identifié 169 emojis sentimentaux[7] à partir de leur représentation (*i.e. son triplet de scores de polarité décrit ci-après*), et de l'*Emoji Sentiment Ranking* (ESR) (Novak *et al.*, 2015). L'ESR fournit les scores de polarité négative, neutre et positive pour 751 emojis à partir d'une annotation manuelle par 83 annotateurs de 1,6 million de tweets en contexte effectué pour 13 langues européennes. Ainsi, l'emoji 😡 qui est représenté par le triplet {négatif ; neutre ; positif} suivant {0,532 ; 0,108 ; 0,360}, est porteur de sentiment. Ce qui n'est pas le cas pour l'emoji ✨ ({0,052 ; 0,545 ; 0,403}) dont la valeur neutre est supérieure aux autres. Bien entendu cet emoji ✨ pourrait être porteur de sentiment dans certains contextes, mais l'utilisation de l'ESR permet d'obtenir la valeur globale moyenne issue de nombreux contextes d'apparition pour chaque emoji.

	Corpus étendu	**Corpus dédié**
Nombre de phrases	88882	9700
Mots	607776	69930
Emojis	148928	18384
Emojis distincts	1070	164
Taux d'emojis sentimentaux	43.34%	100%
Nombre moyen d'emojis par phrase	1,68	1,90
Longueur moyenne des phrases	6 mots	7 mots
Phrases positives ssth*	5832	1014
Phrases négatives ssth*	0	0
Humeurs distinctes utilisées	38	38

Tableau 2 : Caractéristiques des deux corpus utilisés (l'un dédié aux emojis sentimentaux, l'autre étendu à tous les emojis). ssth = valeurs prédites avec SentiStrength

3.1. *Ensemble des caractéristiques*

Nous avons utilisé une représentation vectorielle des phrases du corpus. Cette représentation vectorielle peut varier de dimension en fonction des caractéristiques

6. Modèle d'OpenNLP de découpage en phrases disponible ici : `http://opennlp.sourceforge.net/models-1.5/`
7. `https://gguibon.github.io/coria2017_data.html`

considérées. Nous avons évalué les performances des classifieurs en testant plusieurs combinaisons de caractéristiques. L'ensemble des caractéristiques disponibles est le suivant :

Sac de mots/caractères et nombre de mots. Le contenu textuel peut être représenté de deux façons différentes : soit par un sac de mots, soit par un sac de caractères. Le nombre de mots contenus dans la phrase est également ajouté comme caractéristique car il diffère du nombre d'espaces, ces derniers pouvant être répétés entre les mots ou en fin de phrase. Ainsi la phrase "I love you" sera représentée comme {I} {love} {you} en sac de mots, et {I} {l} {o} {v} {e} {y} {o} {u} en sac de caractères.

N-grammes. Aux sacs de mots sont ajoutés les associations de mots/caractères. Des bi-grammes de la phrase "I love you" donneront donc {I+love} {love+you} en sac de mots, et {I+l} {l+o} {o+v} {v+e} {e+y} {y+o} {o+u} en sac de caractères. Nous nous sommes restreint à un maximum de 5-grammes puisque au-delà l'amélioration des performances stagne, et que cela permet une amélioration de toutes les métriques utilisées (Tableau 5).

L'humeur. Nous avons accès à l'humeur de l'utilisateur. L'utilisateur peut en effet sélectionner un état d'humeur et le changer quand il le souhaite. Ainsi pour chaque message, et donc pour chaque phrase, nous avons accès à l'humeur affichée par l'utilisateur lors de la rédaction du message. Il y a 38 humeurs, chacune étant représentée par un emoji dans l'interface de sélection de l'humeur (bien, très bien, seul, triste, et ainsi de suite).

Polarité de la phrase. Pour chaque phrase, nous prédisons sa polarité afin d'obtenir de nouvelles caractéristiques représentatives du sentiment que peut traduire le ou les emojis porteurs de sentiment. Pour ce faire nous avons utilisé le logiciel SentiStrength[8] (Thelwall *et al.*, 2010) avec le modèle pré-entrainé sur des commentaires de MySpace et des tweets. Il exploite des ressources lexicales propres à la grammaire, à l'orthographe de la communication en ligne (*slang words*, répétitions de caractères), et à la polarité pour prédire deux scores de polarité dans un texte. Un score positif, et un score négative, tous deux allant d'une échelle de 1 (neutre) à 5 et de -1 (neutre) à -5. SentiStrength (Thelwall *et al.*, 2010) est un des modèles d'analyse de sentiment les plus adaptés aux données que nous utilisons puisqu'il a été appris sur des messages informels courts non standardisés. A notre connaissance, il n'existe pas de modèle appris sur des messages courts privés, les tweets et commentaires publiques constituant actuellement les données d'apprentissage habituelles pour cette tâche. Ainsi la prédiction de la polarité pourrait être effectuée par d'autres outils tels que Echo (Hamdan *et al.*, 2015) (SemEval2013) ou encore le modèle de Stanford (Socher *et al.*, 2013) qui nécessiterait de transformer le corpus en corpus arboré.

Présence d'un point d'interrogation/exclamation. Ces deux éléments de ponctuation sont souvent utilisés à des fins d'accentuation, d'autant plus lors de messages instantanés puisque leur utilisation n'est en rien obligatoire. Afin de conserver la consis-

8. http://sentistrength.wlv.ac.uk/

tance de cette caractéristique nous avons explicitement ajouté une caractéristique binaire permettant d'indiquer la présence d'un point d'interrogation dans la phrase ou d'un point d'exclamation. Il convient toutefois de noter que nous n'avons pas pris en compte la répétition de ces signes, cela étant déjà pris en compte dans les modèles d'analyse de sentiment utilisés.

4. Prédiction des emojis

Pour l'apprentissage des modèles nous avons utilisé les *ML-RandomForest* implémentés dans SciKit-Learn[9] avec un total de 20 arbres de décisions sans limiter leur taille. Bien que différents en certains points, chaque modèle a été construit de la manière suivante :

1) Mélange aléatoire des phrases du corpus en question (corpus étendu ou dédié - Tableau 2 -)

2) Extraction des caractéristiques (*features*)

3) Création d'un jeu de test (30%) et d'entrainement (70%)

4) Entrainement d'un classifieur multi-étiquette (*ML-RandomForest*)

5) Prédiction des étiquettes (*i.e.* emojis) sur le jeu de test

6) Evaluation par étiquette (*i.e.* emoji) et évaluation globale du classifieur

4.1. *Prédiction des emojis sentimentaux dans un corpus dédié*

Dans un premier temps nous essayons de prédire les emojis sentimentaux dans les phrases à l'aide d'un corpus réduit en classes possibles et dédié aux 169 emojis sentimentaux ("corpus dédié" au Tableau 2). Ces 169 emojis sentimentaux sont les classes possibles, les autres classes sont donc ignorées.

En appliquant le protocole détaillé en début de section 4 nous obtenons les résultats visibles dans le tableau 3. L'entrainement est effectué sur 70% du corpus dédié aux emojis sentimentaux (6790 phrases) et le test effectué sur les 30% restants (2910 phrases), le tout en 3 itérations afin d'obtenir les moyennes visibles au tableau 3.

Méthode d'évaluation. Dans ces résultats il convient de noter que l'exactitude (*accuracy*) correspond ici à la pertinence moyenne des emojis et non celle du jeu d'étiquettes complet (*Powerset*). Ainsi, si une phrase est étiquetée 😍 et 😘, chaque emoji est considéré séparemment. Nous ne faisons donc pas d'évaluation par *PowerSet*. Par exemple, la phrase donnée précédemment (Tableau 1) pour laquelle seul l'emoji 😭 a été prédit alors que deux emojis étaient attendus, ne verra que l'exactitude de l'emoji correctement prédit augmenter.

9. http://scikit-learn.org/

	Exactitude	Précision	Rappel	Moyenne harmonique
Caractéristiques	**Lemmatisation*, 1 à 5-grammes, tf-idf**			
Sac de mots	48,17	89,65	50,93	64,04
Sac de caractères	55,12	92,19	57,63	70,13
Caractéristiques	**Lemmatisation*, 1 à 5-grammes, nombre de mots tf-idf, score négatif, score positif, exclamation, interrogation**			
Sac de mots	49,94	90,65	52,49	65,63
Sac de caractères	55,31	92,3	57,83	70,31
Caractéristiques	**Lemmatisation*, 1 à 5-grammes, humeur, nombre de mots tf-idf, score négatif, score positif, exclamation, interrogation**			
Sac de mots	54,31	88,63	58,13	69,56
Sac de caractères	**60,28**	**94,30**	**62,52**	**74,49**
Caractéristiques	**Lemmatisation*, humeur, nombre de mots tf-idf, score négatif, score positif, exclamation, interrogation**			
Sac de mots	54,56	89,13	58,24	69,76
Sac de caractères	59,14	93,78	61,49	73,61
Caractéristiques	**Lemmatisation*, 5-grammes, humeur, nombre de mots tf-idf, score négatif, score positif, exclamation, interrogation**			
Sac de mots	23,97	79,44	25,74	37,24
Sac de caractères	57,22	93,95	59,37	72,00

Tableau 3 : Moyennes des performances de prédiction d'emojis sentimentaux avec *ML-Random Forest*. *La lemmatisation est ignorée pour les sacs de caractères.

L'exactitude affichée dans nos résultats correspond donc à la moyenne des exactitudes de chaque emoji, pondérées par leur fréquence dans le corpus de test. C'est également le cas pour la précision et le rappel.

Le tableau 3 suit une logique d'incrémentation des caractéristiques, et représente l'approche de multi-étiquetage par adaptation. L'approche par transformation est ici trop coûteuse puisqu'elle revient à mettre en place 164 classifieurs de *Random Forest*. De plus, cette approche est moins performante qu'une approche par adaptation (*ML-RandomForest*) : avec normalisation tf-idf, et 1 à 5 grammes de caractères, les résultats sont de 91,83% contre 92,19% en précision, 57,37% contre 57,63% en rappel, et 69,8% contre 70,13 en f-mesure. Soit légèrement moins efficace que ceux obtenus avec un multi-étiquetage par adaptation montrés dans le tableau 3 pour les mêmes données et les mêmes caractéristiques utilisées.

Les modèles de multi-étiquetage présentés dans le tableau 3 permettent de prédire les emojis d'une phrase avec une f-mesure maximale de 74,49%, pour une bonne précision de 94,30%, mais un rappel de seulement 62,52%. Cette meilleure prédiction est obtenue en considérant les sacs de caractères avec des n-grammes allant de 1 à 5 caractères, l'humeur de l'utilisateur lors de l'écriture du message, le nombre de mots total de la phrase, les scores de polarité négatifs et positifs prédits, la présence d'un

point d'exclamation et celle d'un point d'interrogation. Le tout est normalisé par tf-idf et lemmatisation.

Analyse de l'impact de la polarité. Les résultats montrent qu'une fois les scores de polarités négatives et positives ajoutés comme caractéristiques, les résultats du classifieur n'obtiennent pas un gain notable ; et ce bien qu'elles soient accompagnées du nombre de mots et de la présence ou non d'un point d'exclamation et d'interrogation. Ce résultat tend à montrer que l'analyse de sentiment en utilisant SentiStrength[8] (Thelwall *et al.*, 2010) n'est pas pertinente pour prédire les emojis sentimentaux. Dans la section 4.2, nous proposons d'explorer la prédiction d'emojis dans le corpus étendu (Tableau 2).

Analyse de l'impact des sacs de caractères/mots. Le tableau 3 réfute l'idée d'associer systématiquement un mot avec un emoji. Avec les sacs de mots, les mots présents dans les représentations vectorielles de deux phrases sont comparés, ce qui nécessite qu'ils soient identiques, même après lemmatisation. La lemmatisation est effectuée à l'aide du *WordNetLemmatizer* de NLTK[10]. Avec l'approche par sac de caractères, la langue incorrecte ou non standardisée et les abbréviations vont être considérées de manière implicite. Par exemple le "you" aura au moins un point commun avec son abbréviation "u", ce qui n'est pas le cas dans un sac de mots. Bien entendu, d'autres mots aurons ainsi un point commun avec "you", mais ceci explique probablement en partie la nette hausse de performance lors de l'utilisation de sacs de caractères.

Analyse de l'impact de l'humeur. L'humeur a un impact important sur la qualité de la prédiction avec un gain constant d'environ 4% en f-mesure. Cette donnée est propre au corpus utilisé, et semble particulièrement représentative du sentiment de l'utilisateur traduit ensuite par des emojis. Bien qu'elle soit très influente sur la qualité de la prédiction, l'humeur à elle seule ne peut suffire à prédire les emojis, le contenu textuel et son mode de représentation (*i.e.* sac de caractères/mots) doit y être associé. Les corpus utilisés montrent ainsi que les emojis varient principalement selon le texte et l'humeur.

Caractéristiques discriminantes. Pour analyser l'importance de chaque caractéristique, nous comparons le score d'importance calculé par le *ML-RandomForest* pour chaque test (Tableau 4). Ces scores conférés après entraînement du classifieur confirment l'importance de l'humeur dans la prédiction des emojis. Seules cette caractéristique, et celle du nombre de mots, semblent avoir des scores élevés. Etant basées sur des sacs de caractères, les caractéristiques peuvent paraître souvent difficilement intelligibles (ex :"ways" en 4ème position dans le tableau 4). Pour une approche par sacs de mots avec humeur et polarités, les 5 caractéristiques les plus importantes sont l'humeur, le nombre de mots, le mot "u" (you), le mot "love" et le point (Tableau 3). Ces scores d'importance sont fonction du nombre total de caractéristiques calculées, d'où leur score relativement bas. En effet il convient de préciser qu'en raison de l'utilisation de différents n-grammes, le nombre de caractéristiques peut être très élevé et ainsi aller de 123 caractéristiques avec 1-gramme à 61268 caractéristiques en utilisant

10. http://www.nltk.org/

1 à 5-grams de caractères, humeur, nombre de mots tf-idf, score positif, score négatif, exclamation, interrogation					
Rang d'importance	**1er**	**2ème**	**3ème**	**4ème**	**5ème**
Caractéristique	Humeur	Nombre mots	Espace	"ways "	"u"
Score	0,0602	0,0051	0,0045	0,0031	0,0029
Lemmatisation, 1 à 5-grams de mots, humeur, nombre de mots tf-idf, score positif, score négatif, exclamation, interrogation					
Rang d'importance	**1er**	**2ème**	**3ème**	**4ème**	**5ème**
Caractéristique	Humeur	Nombre mots	"u"	"love"	"."
Score	0,0997	0,0461	0,0141	0,01368	0,01136

Tableau 4 : Les cinq caractéristiques discriminantes calculées par les *RandomForest*. Avec sac de mots ou de caractères.

des n-grammes de 1 à 5 grammes. Ceci impacte directement la taille de la représentation vectorielle de la phrase ainsi que la valeur d'importance attribuée, sans pour autant changer la hiérarchie des caractéristiques.

	Exactitude	**Précision**	**Rappel**	**F-mesure**
Baseline	53,64	92,26	56,02	68,93
1 à 2 grammes	+1,05	+0,003	+1,12	+0,87
1 à 3 grammes	+1,10	-0,16	+1,3	+0,93
1 à 4 grammes	+1,21	-0,27	+1,41	+0,98
1 à 5 grammes	+1,02	+0,04	+1,1	+0,82
1 à 6 grammes	+0,99	-0,14	+1,12	+0,80
Humeur	**+5,59**	**+2,79**	**+5,04**	**+4,74**
Nombre de mots	+0,0	+0,10	-0,04	+0,0
Score positif*	+0,34	+0,08	+0,37	+0,30
Score négatif*	+0,22	+0,16	+0,18	+0,19
Scores positif et négatif	+0,2	+0,01	+0,25	+0,19
Exclamation	+0,02	-0,06	+0,04	+0,01
Interrogation	+0,03	-0,06	+0,04	+0,02

Tableau 5 : Impact empirique des caractéristiques analysées isolément avec uniquement un sac de caractères en guise de baseline (moyennes obtenues pour 3 exécutions). *Scores prédits avec SentiStrength

La grande majorité de ces caractéristiques sont peu discriminantes mais influencent les résultats une fois regroupées. Bien que les résultats visibles dans le tableau 3 montrent que celles-ci ne sont pas déterminantes, nous avons cherché à vérifier cette conclusion ainsi que les scores d'importance par une quantification de l'impact de chacune d'elles. Dans le tableau 4, nous avons analysé le score d'importance des caractéristiques en considérant les caractéristiques ensemble. Nous complétons cette analyse par une étude de l'impact de chaque caractéristique prise isolément (Tableau 5). Par exemple, la ligne de l'humeur représente les performances de classification

en considérant uniquement le sac de caractères et l'humeur. Les résultats sur cette ligne confirment l'importance de cette caractéristique et l'influence qu'elle peut avoir sur la performance de la prédiction. Mais également que les utilisateurs en font une utilisation pertinente pour le classifieur. Le tableau 5 permet ainsi de voir l'impact des caractéristiques isolées. Il s'agit donc d'une approche empirique dans laquelle la *baseline* est uniquement composée d'un sac de caractères.

Nous observons qu'augmenter les n-grammes semble apporter un peu de performance mais augmente énormément le nombre de caractéristiques créées. Ces n-grammes sont appliqués à l'aide du *TfidfVectorizer* de Scikit, et concernent la totalité des caractères y compris les espaces et ponctuations, ce qui explique pourquoi le point se retrouve présent dans le tableau 4. Aussi, encore une fois l'humeur est confirmée comme étant la caractéristique la plus discriminante.

Enfin, il est intéressant de constater que contrairement aux scores d'importance attribués par le classifieur aux caractéristiques utilisées, le nombre de mots ne semble ici n'avoir aucun importance. Le score de polarité positive a seulement un impact minime mais tout de même plus élevé que le score de polarité négative, ce qui s'explique par le fait que ce dernier ne varie jamais. Ceci constitue une limite de l'utilisation de SentiStrength (Thelwall *et al.*, 2010) sur notre corpus.

Dans cette section, nous avons construit un modèle de classification multi-étiquette pour prédire les emojis sentimentaux. Dans la section suivante, nous proposons d'explorer cette même méthode pour prédire tout type d'emoji.

4.2. *Prédiction des emojis dans un corpus étendu*

Pour construire un modèle de prédiction d'emojis de tout type, nous avons utilisé la même approche sur un corpus uniquement composé de phrases avec emojis (Tableau 2), qu'ils soient sentimentaux ou non. Ce corpus est de taille plus grande mais le nombre classes possibles augmente également. Les 88882 phrases peuvent désormais appartenir à 1070 classes (emojis) différentes. Notre corpus ne contenant pas tous les emojis standards, et se limitant à ceux réellement utilisés et présents dans notre jeu de données, nous avons 1070 emojis au total et non les 2389 emojis d'Unicode. Aussi, les 164 emojis sentimentaux identifiés représentent 43.34% des emojis utilisés, les 906 emojis restants représentant 56.66% des utilisations d'emojis. La représentation des classes n'est donc pas équilibrée dans le corpus. Compte tenu des limites d'obtention de données réelles, et du déséquilibre de l'utilisation des emojis (Pavalanathan et Eisenstein, 2015) déjà visible sur Twitter par le biais de l'EmojiTracker[11], il est très difficile d'équilibrer les classes utilisées.

Comme précédemment le tableau 6 montre les moyennes de trois exécutions aléatoires entrainées sur 70% du corpus (soit 62217 phrases) et testées sur les 30% restant (soit 26665 phrases). Toutes les étiquettes d'humeur sont présentes en entraînement

11. `http://www.emojitracker.com/`

	Exactitude	Précision	Rappel	F-mesure
Caractéristiques	**1 à 5-grammes, tf-idf**			
Sac de caractères	60,44	94,78	62,07	73,76
Caractéristiques	**1 à 5-grammes, nombre de mots tf-idf, score négatif, score positif, exclamation, interrogation**			
Sac de caractères	60,48	94,76	62,11	73,79
Caractéristiques	**1 à 5-grammes, humeur, nombre de mots tf-idf, score négatif, score positif, exclamation, interrogation**			
Sac de caractères	63,91	94,66	65,86	76,63
Caractéristiques	**humeur, nombre de mots tf-idf, score négatif, score positif, exclamation, interrogation**			
Sac de caractères	61,31	94,28	63,28	74,79
Caractéristiques	**5-grams de lettres, humeur, nombre de mots tf-idf, score négatif, score positif, exclamation, interrogation**			
Sac de caractères	62,30	95,94	63,62	75,32

Tableau 6 : Moyennes des prédictions d'emojis sentimentaux avec Random Forest et sac de caractères

et en test. Les résultats globaux présentent une légère amélioration mais restent similaires à ceux obtenus sur le corpus dédié aux emojis sentimentaux. Il y a toujours un fort impact de l'humeur dans la prédiction des emojis. En effet, si l'on en croit les travaux réalisés sur les emojis de Twitter, avec notamment l'EmojiTracker[11] et l'EmojiSentimentRanking (Novak *et al.*, 2015), ainsi que l'importance (43.34%) des emojis sentimentaux dans notre corpus, les emojis les plus utilisés sont ceux porteurs de sentiment. Il n'est donc pas étonnant de retrouver un impact fort de l'humeur. Impact confirmé par les scores des caractéristiques visibles dans le tableau 7.

1 à 5-grams de caractères, humeur, nombre de mots tf-idf, score positif, score négatif, exclamation, interrogation					
Rang d'importance	**1er**	**2ème**	**3ème**	**4ème**	**5ème**
Caractéristique	Humeur	"lex "	"alex"	" al"	"x "
Score	0,0515	0,0059	0,0058	0,0058	0,0058
Lemmatisation, 1 à 5-grams de mots, humeur, nombre de mots tf-idf, score positif, score négatif, exclamation, interrogation					
Rang d'importance	**1er**	**2ème**	**3ème**	**4ème**	**5ème**
Caractéristique	Humeur	"alex"	Nombre mots	"simply"	"pugla*"
Score	0,0688	0,0408	0,0267	0,01481	0,0049

Tableau 7 : Les cinq caractéristiques discriminantes calculées par les *RandomForest*. Avec sac de mots ou de caractères. Pour ce dernier, les scores stagnent à 0.0058 au delà de la 5^{e} position)

Pour confirmer cela nous comparons les performances du même modèle sur les deux ensembles d'emojis : ceux porteurs de sentiment et les autres (Tableau 8). Que

	Exactitude	Précision	Rappel	F-mesure
Caractéristiques	**1 à 5-grammes, nombre de mots tf-idf, score négatif, score positif, exclamation, interrogation**			
Emojis sentimentaux	59,69	94,12	61,71	73,69
Emojis autres	61,69	95,49	62,96	74,32
Caractéristiques	**1 à 5-grammes, humeur, nombre de mots tf-idf, score négatif, score positif, exclamation, interrogation**			
Emojis sentimentaux	63,22	93,33	65,93	76,60
Emojis autres	64,70	95,31	66,26	76,84

Tableau 8 : Comparaison des performances d'un même modèle général sur les deux jeux d'étiquettes (moyennes de 3 exécutions au découpages aléatoires mais comparables)

ce soit avec ou sans l'intégration de l'humeur de l'utilisateur, la prédiction est plus performante sur les autres emojis, ceux qui ne font pas partie des 169 emojis sentimentaux.

Tous ces résultats sur le corpus étendu, et l'écart ténu entre les performances infirme l'hypothèse selon laquelle, parce qu'ils sont vecteurs d'émotions, les emojis sentimentaux seraient mieux prédits à l'aide d'un modèle de classification dédié. Par exemple, l'emoji neutre (selon les polarités de l'ESR) 😴 obtient une f-mesure de 81% et un taux d'exactitude de 68%.

5. Conclusion

Nous avons présenté une approche de classification multi-étiquettes à l'aide de *ML-RandomForest* pour prédire les emojis possibles dans un message. Nous nous sommes limité au niveau de la phrase pour recommander les emojis à l'aide d'un indicateur d'humeur propre à notre corpus et de valeurs de polarité prédites. Nous avons analysé et quantifié l'impact des caractéristiques sur la prédiction des emojis, concluant ainsi qu'il est préférable d'utiliser des sacs de caractères et non de mots.

Afin de confirmer ou d'infirmer le besoin de prédire les emojis porteurs de sentiments à l'aide d'un modèle de classification dédié aux sentiments, nous avons comparé les performances de la prédiction sur deux corpus : l'un dédié aux emojis sentimentaux, l'autre général et comprenant tout type d'emojis. Les résultats montrent qu'un modèle orienté vers le sentiment ne semble pas améliorer uniquement la prédiction des emojis sentimentaux. C'est le cas de notre modèle dédié aux sentiment que nous utilisons. Nous obtenons finalement un modèle de prédiction performant qui maximise la précision, et permet de prédire un grand ensemble d'emojis (1070) dans un message de langue non standard.

Il convient toutefois de souligner que le modèle d'analyse de sentiment utilisé n'est pas tout à fait adapté aux données, puisqu'il n'existe pas de tel modèle à notre connaissance. En utilisant SentiStrength (Thelwall *et al.*, 2010), nous avons utilisé un classifieur appris sur des données plus standardisées car issues de tweets et de commentaires MySpace. Tous deux consistant des messageries sociales publiques, tandis que nous utilisons des données privées qui ne sont pas soumises à des limitations de taille, ou à l'obligation de rendre son message compréhensible par le plus grand nombre. La preuve en est de l'attribution d'une valeur négative à aucune phrase de nos corpus utilisés (Tableau 2), même si les valeurs positives prédites compensent ce déficit.

Pour la suite de nos travaux, il serait donc nécessaire de pallier à ce déficit en utilisant également d'autres méthodes d'analyse de sentiment basées sur des caractéristiques lexicales (Hamdan *et al.*, 2015). Ensuite nous pourrons étendre le contexte pris en compte à celui de la conversation en obtenant ainsi l'historique des messages, ce qui permettrait également d'orienter la recommandation en fonction du profil de l'utilisateur et des emojis qui y sont associés (Barbera, 2016).

6. Bibliographie

Alm C. O., Roth D., Sproat R., « Emotions from text : machine learning for text-based emotion prediction », *Proceedings of the conference on human language technology and empirical methods in natural language processing*, Association for Computational Linguistics, p. 579-586, 2005.

Avazpour I., Pitakrat T., Grunske L., Grundy J., « Dimensions and metrics for evaluating recommendation systems », *Recommendation systems in software engineering*, Springer, p. 245-273, 2014.

Baldridge J., « The opennlp project », *URL : https ://opennlp.apache.org/,(accessed 2 February 2012)*, 2005.

Barbera P., « Less is more ? How demographic sample weights can improve public opinion estimates based on Twitter data », 2016.

Breiman L., « Random forests », *Machine learning*, vol. 45, n° 1, p. 5-32, 2001.

Chaffar S., Inkpen D., « Using a heterogeneous dataset for emotion analysis in text », *Canadian Conference on Artificial Intelligence*, Springer, p. 62-67, 2011.

De Oliveira M. G., Ciarelli P. M., Oliveira E., « Recommendation of programming activities by multi-label classification for a formative assessment of students », *Expert Systems with Applications*, vol. 40, n° 16, p. 6641-6651, 2013.

Eisner B., Rocktäschel T., Augenstein I., Bošnjak M., Riedel S., « emoji2vec : Learning Emoji Representations from their Description », *arXiv preprint arXiv :1609.08359*, 2016.

Ekman P., « Facial expression and emotion. », *American psychologist*, vol. 48, n° 4, p. 384, 1993.

Hamdan H., Bellot P., Bechet F., « Sentiment lexicon-based features for sentiment analysis in short text », *In Proceeding of the 16th International Conference on Intelligent Text Processing and Computational Linguistics*, 2015.

Huang S., Peng W., Li J., Lee D., « Sentiment and topic analysis on social media : a multi-task multi-label classification approach », *Proceedings of the 5th annual ACM web science conference*, ACM, p. 172-181, 2013.

Kelly R., Watts L., « Characterising the inventive appropriation of emoji as relationally meaningful in mediated close personal relationships », 2015.

Lauser B., Hotho A., « Automatic multi-label subject indexing in a multilingual environment », *International Conference on Theory and Practice of Digital Libraries*, Springer, p. 140-151, 2003.

Lenhart A., Smith A., Anderson M., Duggan M., Perrin A., « Teens, technology and friendships », 2015.

Li W., Xu H., « Text-based emotion classification using emotion cause extraction », *Expert Systems with Applications*, vol. 41, n° 4, p. 1742-1749, 2014.

Miller H., Thebault-Spieker J., Chang S., Johnson I., Terveen L., Hecht B., « Blissfully happy" or "ready to fight" : Varying Interpretations of Emoji », *Proceedings of ICWSM 2016*, 2016.

Mishne G. *et al.*, « Experiments with mood classification in blog posts », *Proceedings of ACM SIGIR 2005 workshop on stylistic analysis of text for information access*, vol. 19, Citeseer, p. 321-327, 2005.

Nair-Benrekia N.-Y., Classification interactive multi-label pour l'aide à l'organisation personnalisée des données, PhD thesis, Université de Nantes, 2015.

Novak P. K., Smailović J., Sluban B., Mozetič I., « Sentiment of emojis », *PloS one*, vol. 10, n° 12, p. e0144296, 2015.

Pavalanathan U., Eisenstein J., « Emoticons vs. emojis on Twitter : A causal inference approach », *arXiv preprint arXiv :1510.08480*, 2015.

Peirce C. S., « Logic as semiotic : The theory of signs », 1902.

Rubin T. N., Chambers A., Smyth P., Steyvers M., « Statistical topic models for multi-label document classification », *Machine learning*, vol. 88, n° 1-2, p. 157-208, 2012.

Socher R., Perelygin A., Wu J. Y., Chuang J., Manning C. D., Ng A. Y., Potts C. *et al.*, « Recursive deep models for semantic compositionality over a sentiment treebank », *Proceedings of the conference on empirical methods in natural language processing (EMNLP)*, vol. 1631, Citeseer, p. 1642, 2013.

Strobl C., Boulesteix A.-L., Kneib T., Augustin T., Zeileis A., « Conditional variable importance for random forests », *BMC bioinformatics*, vol. 9, n° 1, p. 307, 2008.

Team E. R., « Emoji_Report », 2015.

Thelwall M., Buckley K., Paltoglou G., Cai D., Kappas A., « Sentiment strength detection in short informal text », *Journal of the American Society for Information Science and Technology*, vol. 61, n° 12, p. 2544-2558, 2010.

Trohidis K., Tsoumakas G., Kalliris G., Vlahavas I. P., « Multi-Label Classification of Music into Emotions. », *ISMIR*, vol. 8, p. 325-330, 2008.

Tsoumakas G., Katakis I., « Multi-label classification : An overview », *International Journal of Data Warehousing and Mining*, 2006.

Xie R., Liu Z., Yan R., Sun M., « Neural Emoji Recommendation in Dialogue Systems », *arXiv preprint arXiv :1612.04609*, 2016.

Zhang M.-L., Zhou Z.-H., « A review on multi-label learning algorithms », *IEEE transactions on knowledge and data engineering*, vol. 26, n° 8, p. 1819-1837, 2014.