

Soutenance alternance M2

Triage automatique de mails



30/09/2022
Maxime POULAIN

Soutenance M2 – Triage automatique de mails

Sommaire

- **Présentation Verlingue**
- **Contexte du projet**
- **Développement technique**
 - La boîte mail VC
 - Structure de la solution
 - Premier algorithme : entraînement du modèle
 - Deuxième algorithme : routage des mails
 - Dashboard et résultats obtenus
- **Conclusion**

Soutenance M2 – Triage automatique de mails

Présentation Verlingue

- Entreprise familiale créé en 1933 à Quimper.
- Courtier en assurances spécialisé dans la protection des entreprises
 - **Obj** : Servir d'interlocuteur entre les entreprises et les assureurs afin de construire, négocier et gérer des programmes d'assurances sur mesure.



- Groupe Adelaïde :
 - **Verlingue** : courtage en assurances auprès des entreprise
 - **Génération** : spécialisé dans la gestion des prestations de santé et de prévoyances
 - **Cocoon** : courtage en assurance auprès des particuliers

Soutenance M2 – Triage automatique de mails

Présentation Verlingue

- DSI commune à l'ensemble du groupe Adelaïde
- Équipe DataFacory
 - Jean-Marie L'HOTELLIER : responsable du domaine BI et mon tuteur
 - Marc ALLAIN : concepteur développeur BI
 - + 3 autres membres
- Besoin sur un projet IA → recrutement d'un alternant

Soutenance M2 – Triage automatique de mails

Contexte du projet

- Verlingue Connect:
 - Service qui fournit des solutions d'assurances aux petites et moyennes entreprises
 - Deux branches d'activités spécifiques
 - ADP : Assurances de personnes
 - IARD : Incendie, Accident et Risques Divers

- Boîte mails générique:
 - Arborescence partagé par la dizaine de collaborateurs
 - Traitement des mails à la volée

⇒ **Perte de temps importante dans la journée !**

Soutenance M2 – Triage automatique de mails

Contexte du projet

- Solution de triage automatique de mails
 - Envoyer les mails qui arrivent dans la boîte de réception dans les dossiers des collaborateurs
- **Mélusine:**
 - Bibliothèque python open-source
 - Développée par la MAIF et Quantmetry
 - Obj : Simplifier le traitement des mails écrits en français

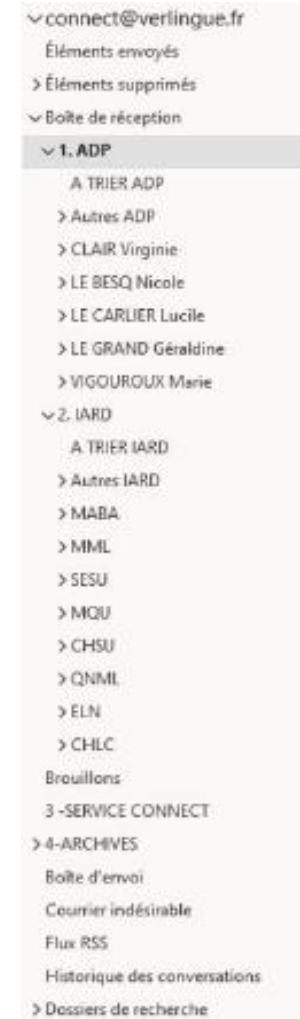


Soutenance M2 – Triage automatique de mails

Développement technique > La boîte mail VC

- Réorganisation de la boîte mail
 - Deux équipes avec une organisation différentes de leurs dossiers
 - Pas au même niveau de l'arborescence
 - Différences au sein des équipes aussi
 - Dossiers en dehors de la boîte de réception
 - Présence de « / » dans les noms des dossiers

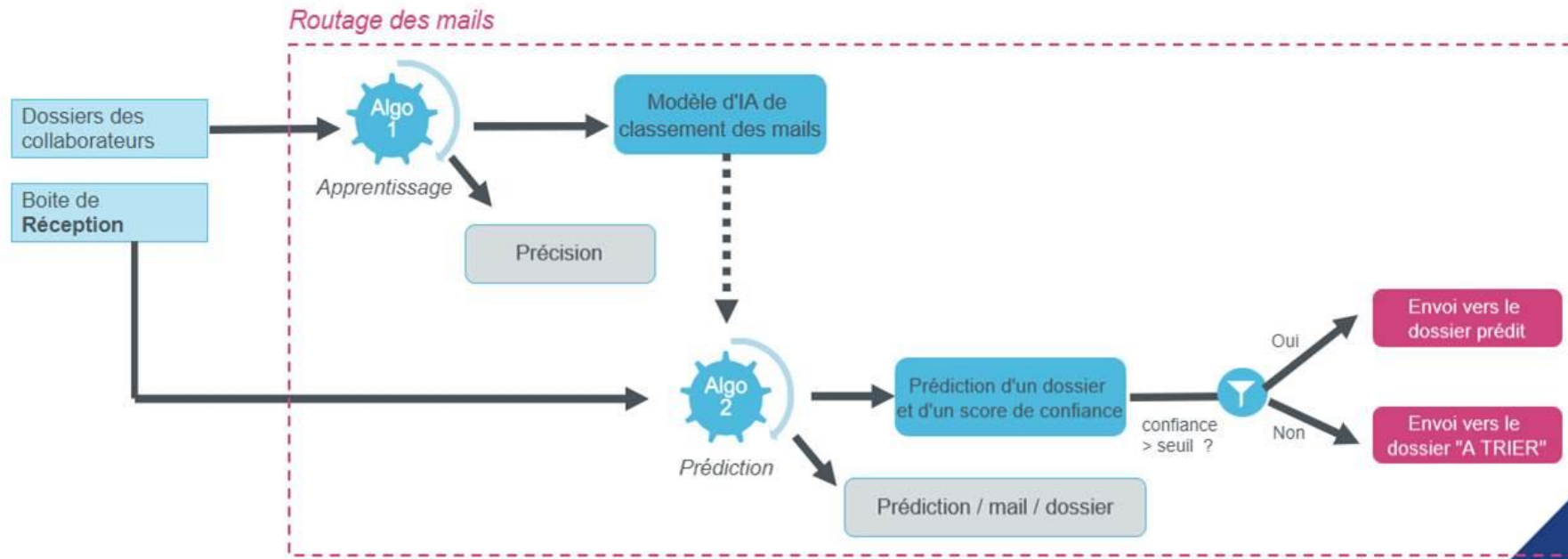
⇒ Changements dans l'arborescence



Soutenance M2 – Triage automatique de mails

Développement technique > Structure de la solution

- Solution basée sur deux algorithmes



Soutenance M2 – Triage automatique de mails

Développement technique > Premier algorithme : entraînement du modèle d'IA

■ 1. Extraction des mails de la boîte

- Obj : Avoir un nombre de mails par collaborateur proche dans le jeu de données

$$Nb\ de\ mails = Objectif\ de\ mails\ par\ collab.* \frac{Nb\ mails\ dans\ ce\ dossier}{Nb\ mails\ total\ du\ collab.}$$

- Récupération des mails et de leurs caractéristiques

Index	message id	body	header	date	from	to	attachment	target
0	<8219607...	IND_DCM No... IND_DMA No...	DWH ...	2022-08-21...	no-reply@verlingue.fr	['DSIDataFactory@netdmx.com...	None	Mail automati...
1	<2b3258d...	Direction ... Quimper	RE7_...	2021-10-06...	DSI-suivi-product...	['DSIDataFactory@netdmx.com...	None	Mail automati...
2	<80e0207...	[Une image...	Comm...	2022-04-07...	bernard.delaporta...	['VerlingueAix@netdmx.com', ...	['image0...	Actualité/RH- Paye
3	<A34F388...	https://i...	Guid...	2021-11-05...	juliette.varet@ve...	['thomas.raher@verlingue.fr...	['Matiné...	Actualité/RH- Paye

Soutenance M2 – Triage automatique de mails

Développement technique > Premier algorithme : entraînement du modèle d'IA

- 2. Traitement des caractéristiques des mails
- Premier pipeline : préparer les données
 - Début de la conversation par un transfert, une réponse ?
 - Nettoyage du corps et de l'objet du dernier mail

Index	target	is begin by transfer	is answer	is transfer	structured historic	structured body	last body	clean body	clean header
0	Mail ...	False	False	False	[{'text': "ATT...	[{'meta': {'...	ATTENTION : email e...	attention : email ...	alarmes majeur sur...
1	Mail ...	False	False	False	[{'text': "\r ...	[{'meta': {'...	Les heures sont exp...	les heures sont ex...	fin alarme sur sit...
2	Actua...	False	False	False	[{'text': '[Un...	[{'meta': {'...	[Une image contenan...	[une image contena...	welcome day : sur ...
3	Actua...	False	False	True	[{'text': 'Bon...	[{'meta': {'...	Bonjour, Pour rapp...	bonjour, pour rapp...	important pour les...

Soutenance M2 – Triage automatique de mails

Développement technique > Premier algorithme : entraînement du modèle d'IA

■ 2. Traitement des caractéristiques des mails

■ Deuxième pipeline : outils de NLP:

- Phraser : regroupe les expressions communes (ex : bulletin salaire => bulletin_salaire)
- Tokenizer : séparer le texte en liste de mots
- Stemmer : réduire les mots à leur racine

Index	structured historic	structured body	last body	clean body	clean header	clean text	tokens	stemmed tokens
0	{'text': "Bon... [{"meta': {'...	Bonjou...	bonjour ...	pot de depart	pot de depart bonjour a ...	['pot', 'depart', 'bonjour'...	['pot', 'depart', 'bonjour'...	
1	{'text': "ATT... [{"meta': {'...	ATTENT...	attentio...	confirmati...	confirmation webinar bus...	['confirmation', 'webinar', ...	['confirm', 'webinar', 'bus...	
2	{'text': '[ci... [{"meta': {'...	[cid:i...	[cid:ima...	ils nous o...	ils nous ont rejoints ! ...	['ils', 'rejoints', 'cid', ...	['il', 'rejoint', 'cid', 'i...	
3	{'text': 'Bon... [{"meta': {'...	Avant ...	avant ce...	fin de moi...	fin de mois : saisie des...	['fin', 'mois', 'saisie', '...	['fin', 'mois', 'sais', 'fe...	

Soutenance M2 – Triage automatique de mails

Développement technique > Premier algorithme : entraînement du modèle d'IA

■ 2. Traitement des caractéristiques des mails

■ Troisième pipeline : traitement des métadonnées:

- Extension des adresses mails (ex: verlingue.fr)
- Date d'envoi
- Type de la pièce jointe
- Nom expéditeur
- Noms destinataires

=> LabelEncoder pour transformer les données catégorielles en données numériques

Index	20	extension 21	extension 22	extension 23	extension 24	davofweek 0	davofweek 1	davofweek 2	davofweek 3	davofweek 4
0	0	1	0	0	1	0	0	0	0	0
1	0	0	0	1	0	0	1	0	0	0
2	0	1	0	0	0	0	0	0	0	0
3	0	1	0	0	0	0	1	0	0	0
4	0	0	0	0	0	0	0	0	0	0

Soutenance M2 – Triage automatique de mails

Développement technique > Premier algorithme : entraînement du modèle d'IA

- **3. entraînement du modèle d'IA**
- **Embedding** : représentation abstraite des mots dans un vecteur de dimension plus petite (ex : Word2Vec)
=> Entraînement sur les tokens
- Modèle **CNN** disponible dans Mélusine
 - Réseau de neurones convolutionnel
- Données à fournir au modèle de Mélusine
 - Le modèle d'Embedding
 - La concaténation du texte et de l'objet traités
 - Les métadonnées

Soutenance M2 – Triage automatique de mails

Développement technique > Deuxième algorithme : routage des mails

- **1. Connexion et extraction des mails**
 - Connexion à la boîte mail toutes les minutes
- **2. Nettoyage des mails**
 - Réutilisation des pipelines entraînés dans le premier algorithme
- **3. Prédiction de l'IA**
 - Prédiction du modèle d'IA entraîné
 - Sauvegarde de la prédiction

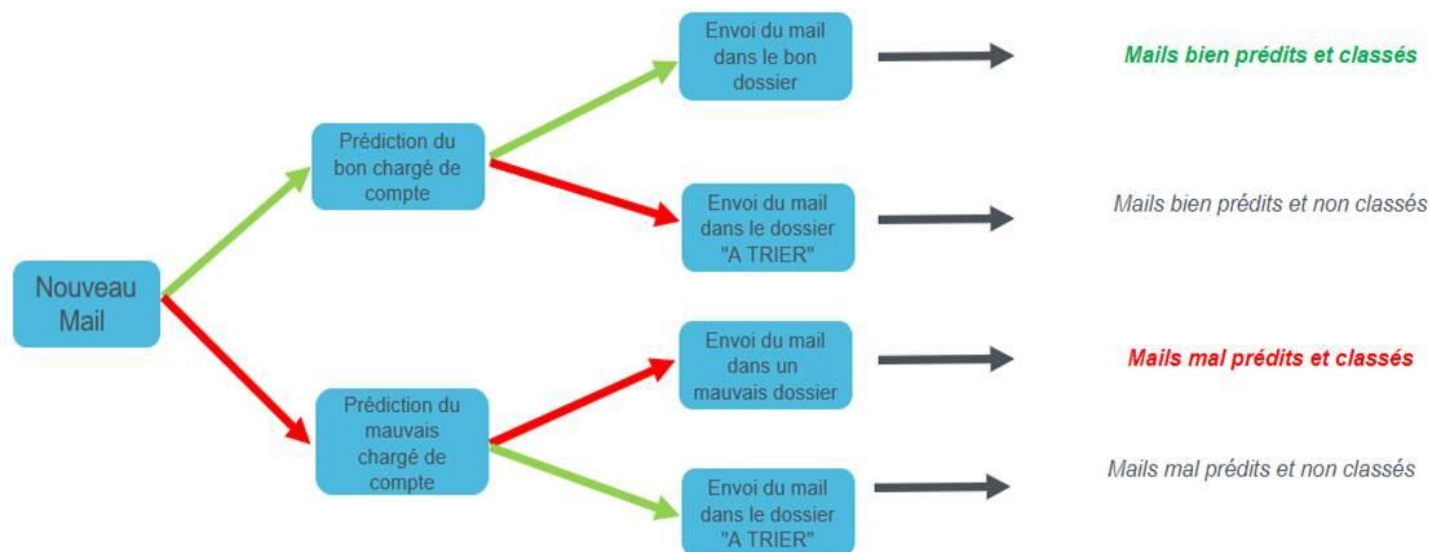
Soutenance M2 – Triage automatique de mails

Développement technique > Deuxième algorithme : routage des mails

■ 4. Déplacement des mails

- Le modèle fournit une prédiction et une probabilité (score de confiance)
=> Utilisation d'un seuil pour limiter les erreurs de routage

- Quatre cas de figure :

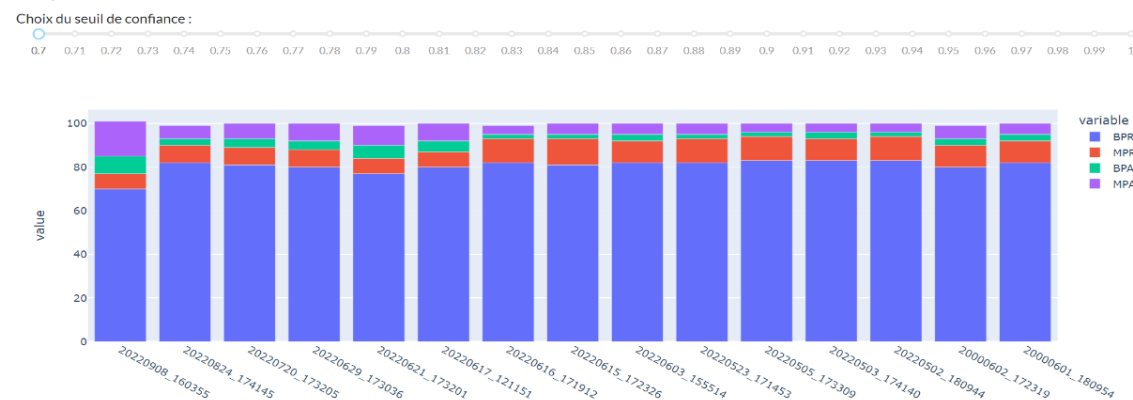


Soutenance M2 – Triage automatique de mails

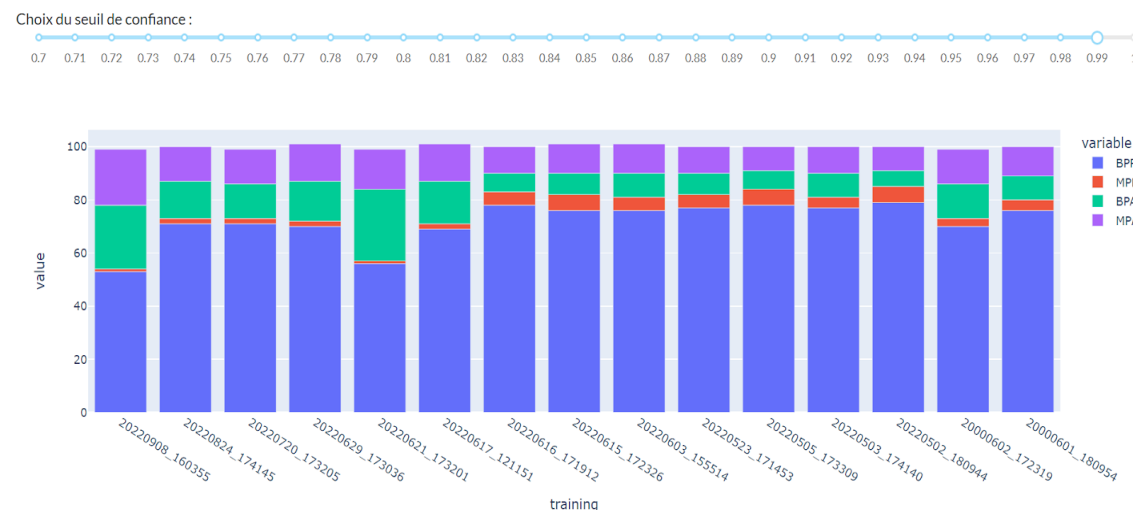
Développement technique > Dashboard et résultats obtenus

■ 1. Comparaison des entraînements des modèles d'IA

Répartition des mails en fonction du seuil de confiance



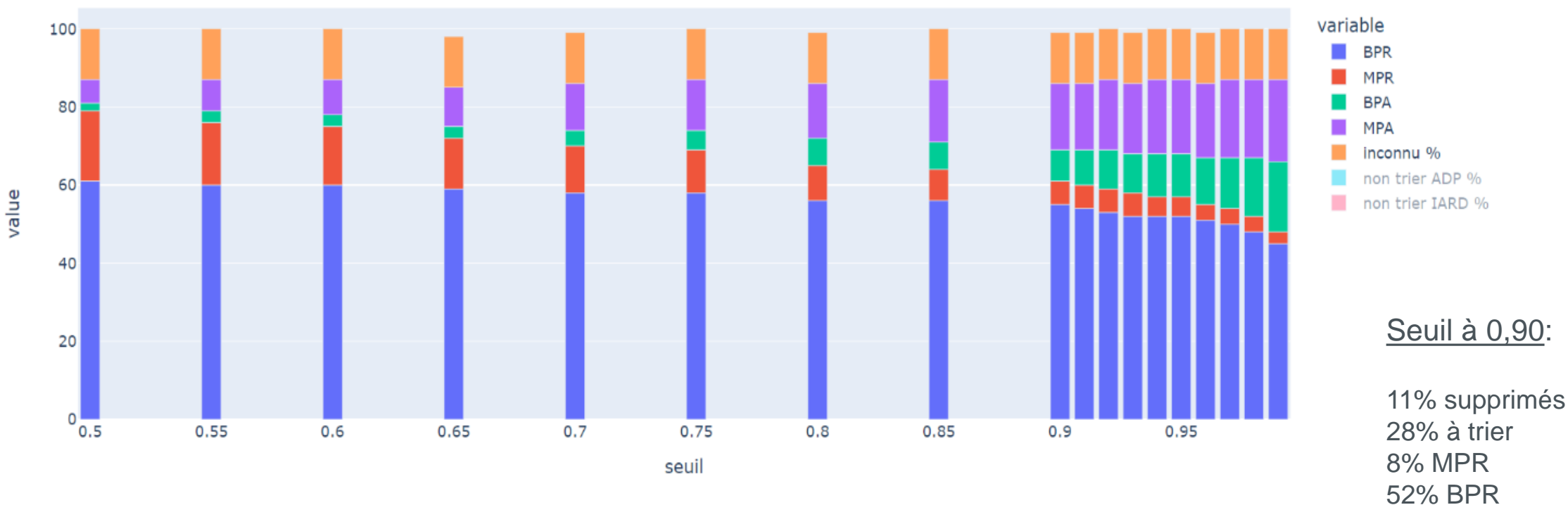
Répartition des mails en fonction du seuil de confiance



Soutenance M2 – Triage automatique de mails

Développement technique > Dashboard et résultats obtenus

■ 2. Résultats des prédictions de l'IA

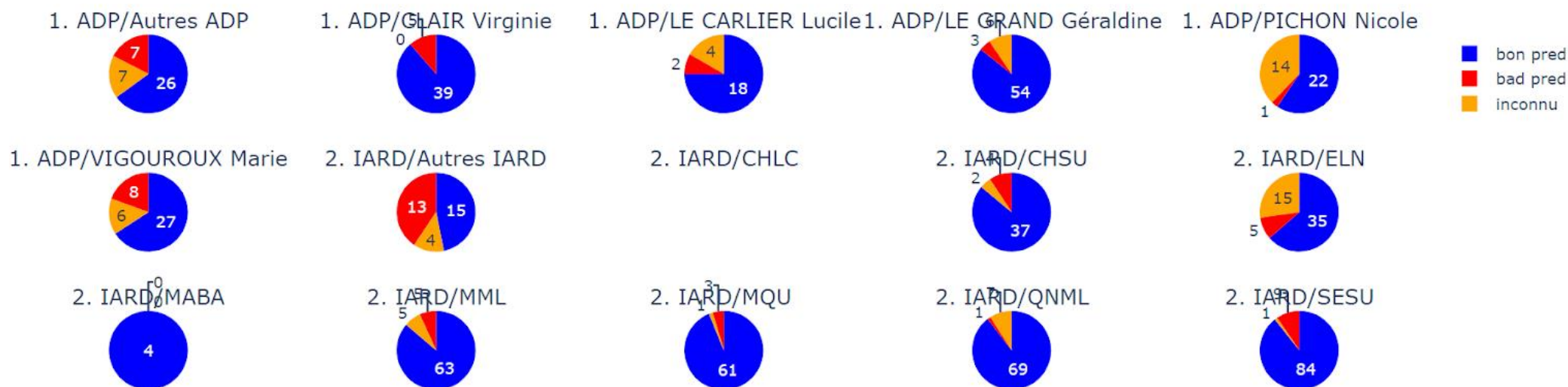


Soutenance M2 – Triage automatique de mails

Développement technique > Dashboard et résultats obtenus

■ 2. Résultats des prédictions de l'IA

Repartition mails reçus par chaque gestionnaire en prenant en compte les seuils utilisés



Soutenance M2 – Triage automatique de mails

Conclusion

- Projet en production depuis mi-juin
 - Métiers content des résultats
 - Demandes pour mise en place sur d'autre boîte mails⇒ Bilan positif
- Perspectives
 - Amélioration des résultats avec de la data augmentation ?
 - ...

Verlingue, simplement

Courtier en assurances spécialisé dans la protection des entreprises, Verlingue apporte à ses clients des solutions simples et performantes pour protéger leurs activités (risques de l'entreprise) et leurs collaborateurs (protection sociale complémentaire).

Avec plus de 1 100 collaborateurs implantés en régions et des partenaires dans plus de 100 pays (membre du réseau WBN), Verlingue conseille, négocie et gère les programmes d'assurances de plus de 3 200 entreprises. Avec une ambition au service de ses clients : rendre simples leurs assurances.

www.verlingue.fr

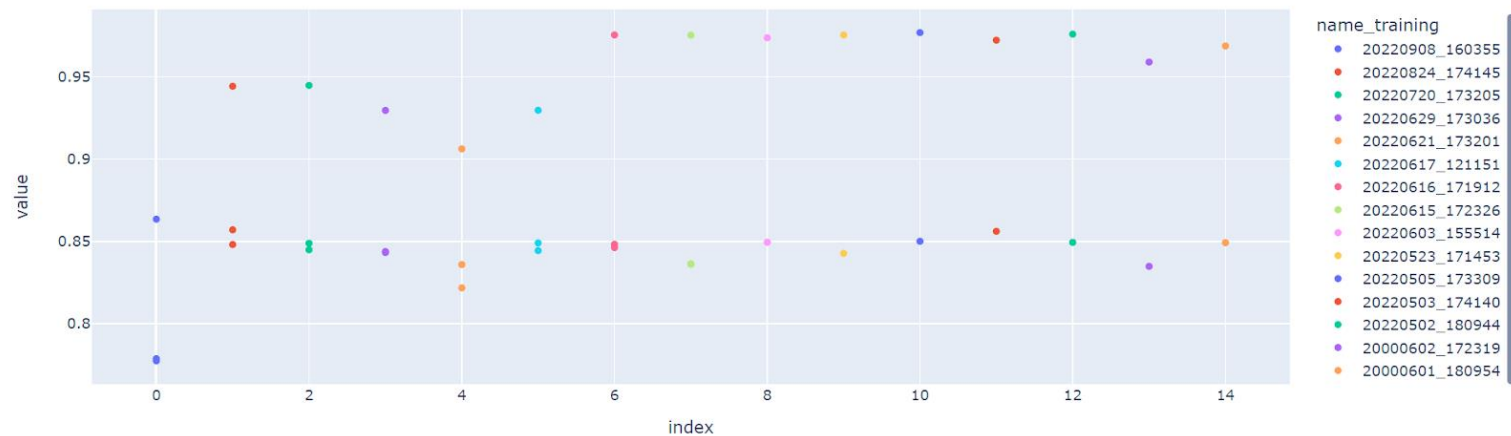
Soutenance M2 – Triage automatique de mails

Développement technique > Dashboard et résultats obtenus

■ 1. Comparaison des entraînements des modèles d'IA

Accuracy

Accuracy de train et de test sur le même schéma



Paramètres d'entraînement

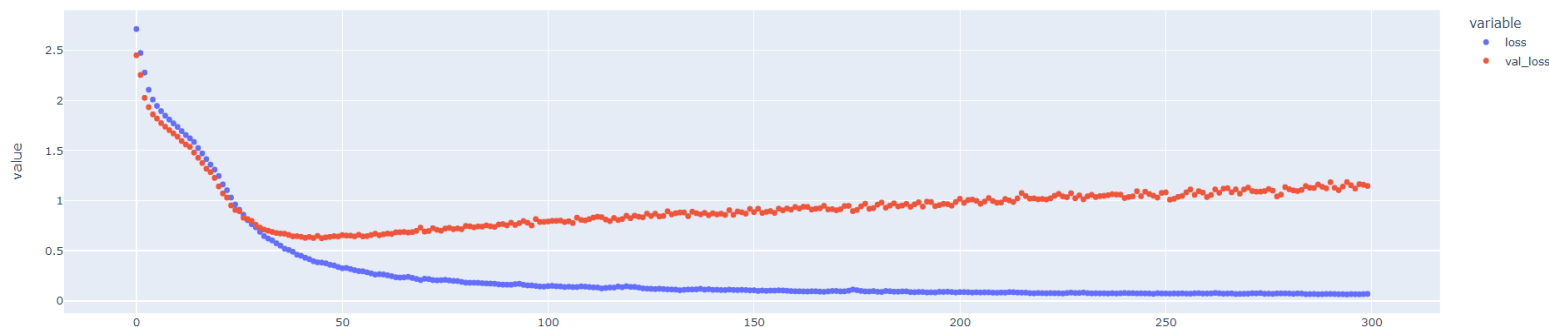
nb epochs	obj mails by target	model name	min emails	folders banned	nom
50	7000	cnn	1	['A TRIER ADP', 'A TRIER IARD']	20220908_160355
50	7000	cnn	1	['A TRIER ADP', 'A TRIER IARD']	20220824_174145
50	7000	cnn	1	['A TRIER ADP', 'A TRIER IARD']	20220720_173205
50	7000	cnn	1	['A TRIER ADP', 'A TRIER IARD']	20220629_173036
50	3000	cnn	1	['A TRIER ADP', 'A TRIER IARD']	20220621_173201
50	7000	cnn	1	['A TRIER ADP', 'A TRIER IARD']	20220617_121151
300	7000	cnn	1	['A TRIER ADP', 'A TRIER IARD']	20220616_171912
300	5000	cnn	1	['A TRIER ADP', 'A TRIER IARD']	20220615_172326
200	5000	cnn	1	['A TRIER ADP', 'A TRIER IARD']	20220603_155514

Soutenance M2 – Triage automatique de mails

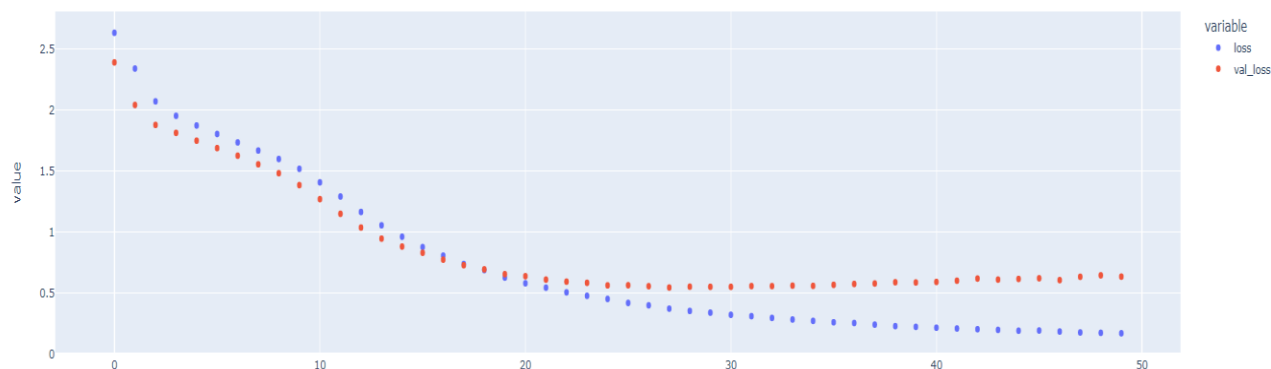
Développement technique > Dashboard et résultats obtenus

■ 1. Comparaison des entraînements des modèles d'IA

Évolution de la perte d'apprentissage



Évolution de la perte d'apprentissage

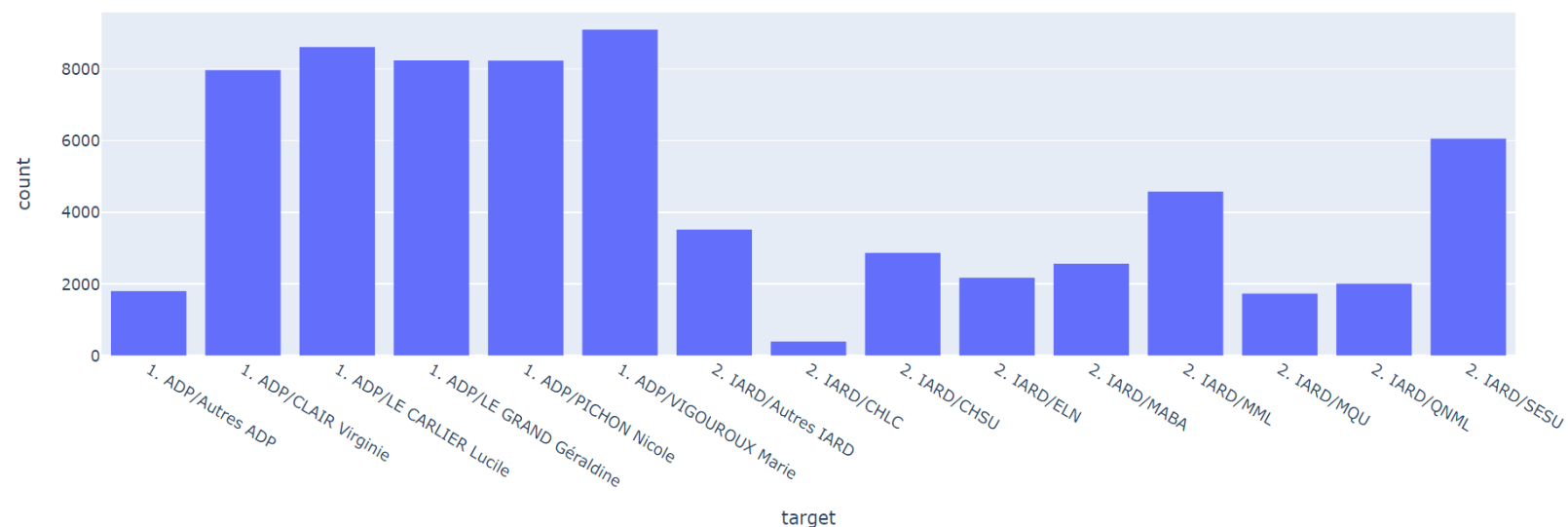


Soutenance M2 – Triage automatique de mails

Développement technique > Dashboard et résultats obtenus

■ 1. Comparaison des entraînements des modèles d'IA

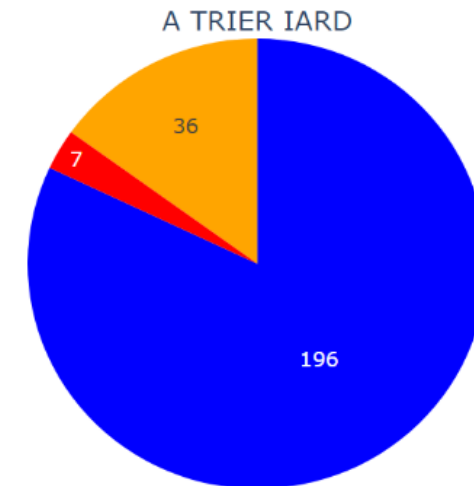
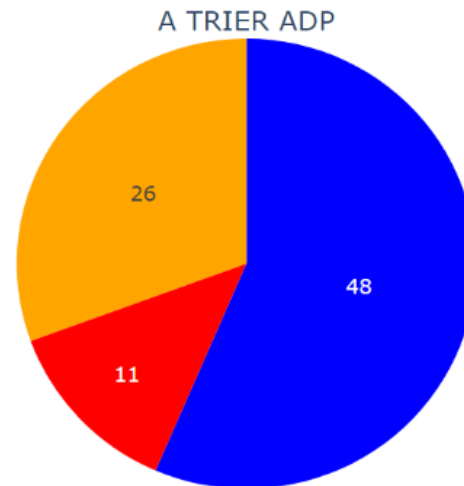
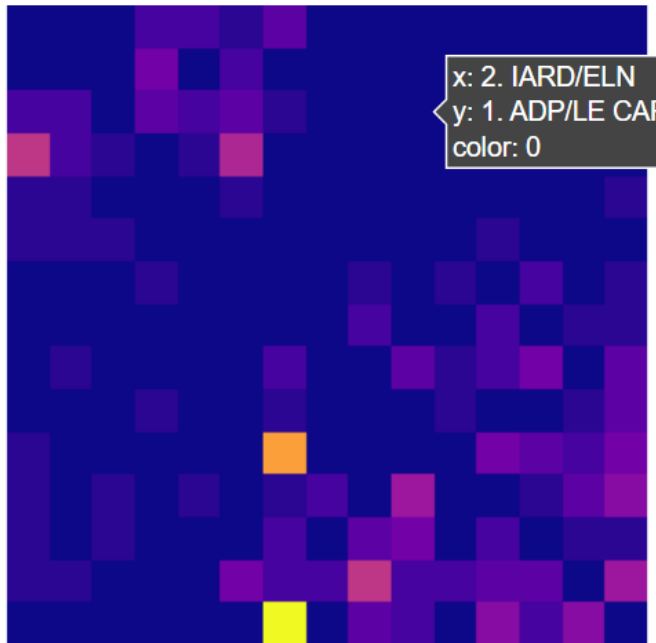
Répartition label



Soutenance M2 – Triage automatique de mails

Bonus

■ 2. Résultats des prédictions de l'IA



■ Ok
■ inconnu
■ confusion ADP/IARD