Une approche hybride pour la détection d'influenceurs dans les médias sociaux

- Projet SOMA (E9202)

soutenu par la commission européenne Eurostars

Namrata Patel[1], Cédric Lopez[1], Ioannis Partalas[1], Frédérique Segond[1]

Viseo Technologies
4 av. Doyen Louis Weil, 38000, Grenoble, France
[prenom].[nom]@viseo.com

VISEO

La détection d'influenceurs dans les médias sociaux

Introduction

Un influenceur, c'est...

« un individu qui par son statut, sa position ou son exposition médiatique peut influencer les comportements de consommation dans un univers donné »

La détection d'influenceurs dans les médias sociaux

Introduction

Un influenceur, c'est...

« un individu qui par son statut, sa position ou son exposition médiatique peut influencer les comportements de consommation dans un univers donné »

La détection d'influenceurs

- Poussée par les intérêts du marketing
- Evaluation automatique de l'influence des individus dans les médias sociaux

Plan

État de l'art

Notre approche

- Les caractéristiques d'influence
- Approche hybride implémentée

Expérimentation

- Les jeux de données
- Évaluation
- Résultats

La détection d'influenceurs dans les médias sociaux

État de l'art

Notre approche

- Les caractéristiques d'influence
- Approche hybride implémentée

Expérimentation

- Les jeux de données
- Évaluation
- Résultats

La détection d'influenceurs

État de l'art

Analyse structurelle du réseau d'interactions

- Graphe représentant les utilisateurs et leurs interactions
- Identification des nœuds les plus importants d'un réseau
- Théorie des graphes : mesures de centralité (i.e. PageRank)

Analyse du contenu textuel

- Identification de marqueurs linguistiques d'influence
- Caractéristiques d'influence tirées d'études en psychologie
- Analyse d'opinions véhiculées par les messages
- Approches par apprentissage automatique

VISEO

La détection d'influenceurs

Notre approche

Constat

Les caractéristiques d'influence tirées d'études en psychologie sont pertinents pour la tâche de détection d'influenceurs. Elles sont extraites des messages d'utilisateurs sous forme de critères statistiques.

La détection d'influenceurs

Notre approche

Constat

Les caractéristiques d'influence tirées d'études en psychologie sont pertinents pour la tâche de détection d'influenceurs. Elles sont extraites des messages d'utilisateurs sous forme de critères statistiques.

Problématique

Ces caractéristiques d'influence sont également présentes dans les messages d'utilisateurs sous forme de critères linguistiques qui sont difficiles à extraire par des approches statistiques.

La détection d'influenceurs

Notre approche

Constat

Les caractéristiques d'influence tirées d'études en psychologie sont pertinents pour la tâche de détection d'influenceurs. Elles sont extraites des messages d'utilisateurs sous forme de critères statistiques.

Problématique

Ces caractéristiques d'influence sont également présentes dans les messages d'utilisateurs sous forme de critères linguistiques qui sont difficiles à extraire par des approches statistiques.

Objectif

Développer une approche qui permet d'évaluer la pertinence de ces critères linguistiques confrontés aux critères statistiques.

VISEO

La détection d'influenceurs

Notre approche

Idée :

- Reprise de travaux antérieurs

La détection d'influenceurs

Notre approche

Idée :

- Reprise de travaux antérieurs
- Sélection de critères linguistiques et statistiques

La détection d'influenceurs

Notre approche

Idée :

- Reprise de travaux antérieurs
- Sélection de critères linguistiques et statistiques
- Développement d'une approche hybride préliminaire :

La détection d'influenceurs

Notre approche

Idée :

- Reprise de travaux antérieurs
- Sélection de critères linguistiques et statistiques
- Développement d'une approche hybride préliminaire :
  - Extraction de critères linguistiques par une approche symbolique

La détection d'influenceurs

Notre approche

Idée :

- Reprise de travaux antérieurs
- Sélection de critères linguistiques et statistiques
- Développement d'une approche hybride préliminaire :
  - Extraction de critères linguistiques par une approche symbolique
  - Prédiction de score d'influence par apprentissage automatique

La détection d'influenceurs

Notre approche

Idée :

- Reprise de travaux antérieurs
- Sélection de critères linguistiques et statistiques
- Développement d'une approche hybride préliminaire :
  - Extraction de critères linguistiques par une approche symbolique
  - Prédiction de score d'influence par apprentissage automatique
- Evaluation de la pertinence des critères linguistiques

La détection d'influenceurs dans les médias sociaux

État de l'art

Notre approche

- Les caractéristiques d'influence
- Approche hybride implémentée

Expérimentation

- Les jeux de données
- Évaluation
- Résultats

Les techniques de l'influence

Robert Cialdini (1984)

Influence situationnelle - études en psychologie

- Réciprocité du don
- Engagement et cohérence (e.g. Inception)
- Preuve par masse
- Autorité
- Rareté (e.g. offre limitée)
- Appréciation et amitié (e.g. influenceurs sur Twitter)

Les caractéristiques de l'influence

Travaux de thèse - Sara Rosenthal (2015)

Descripteurs d'influence pour la détection automatique

- Caractéristiques de l'auteur
- Accord/désaccord
- Déclarations opinées
- Argumentation
- Persuation
- Crédibilité
- Structure du dialogue

Le module de détection SOMA

Etape 1 - Analyse du contenu

Le module de détection SOMA

Etape 1 - Analyse du contenu

Descripteurs d'influence pour la détection automatique

- Accord/désaccord : Définition de règles linguistiques suivant une approche symbolique

Le module de détection SOMA

Etape 1 - Analyse du contenu

Descripteurs d'influence pour la détection automatique

- Accord/désaccord : Définition de règles linguistiques suivant une approche symbolique
- Déclarations opinées : Détection par combinaison d'analyse d'opinions avec reconnaissance d'entités

Le module de détection SOMA

Etape 1 - Analyse du contenu

Descripteurs d'influence pour la détection automatique

- Accord/désaccord : Définition de règles linguistiques suivant une approche symbolique
- Déclarations opinées : Détection par combinaison d'analyse d'opinions avec reconnaissance d'entités
- Argumentation : Module de reconnaissance de termes à partir d'un lexique d'argumentation

Le module de détection SOMA

Etape 1 - Analyse du contenu

Descripteurs d'influence pour la détection automatique

- Accord/désaccord : Définition de règles linguistiques suivant une approche symbolique
- Déclarations opinées : Détection par combinaison d'analyse d'opinions avec reconnaissance d'entités
- Argumentation : Module de reconnaissance de termes à partir d'un lexique d'argumentation
- Structure du dialogue : analyse au niveau de la structure du fil de discussion

Le module de détection SOMA

Etape 1 - Analyse du contenu

Descripteurs d'influence pour la détection automatique

- Accord/désaccord : Définition de règles linguistiques suivant une approche symbolique
- Déclarations opinées : Détection par combinaison d'analyse d'opinions avec reconnaissance d'entités
- Argumentation : Module de reconnaissance de termes à partir d'un lexique d'argumentation
- Structure du dialogue : analyse au niveau de la structure du fil de discussion
- Critères numériques classiques : taille du message, date, etc.

Le module de détection SOMA

Etape 1 - Analyse du contenu

Critères extraits

Critère	Catégorie	Nature	Sortie (type)
isFirstPost ?	non-linguistique	Position d'un message dans un fil	booléen
isSecondPost ?	non-linguistique	Position d'un message dans un fil	booléen
isPenultimateost ?	non-linguistique	Position d'un message dans un fil	booléen
isLatestPost ?	non-linguistique	Position d'un message dans un fil	booléen
sizeOfMessage	non-linguistique	Information quantitative	$0 < x < n$
RegistrationDate	non-linguistique	Date	date
Location of the user	non-linguistique	Emplacement	string
Elongation	linguistique	Style d'écriture	booléen
Uppercase	linguistique	Style d'écriture	booléen
Exclamation	linguistique	Style d'écriture	booléen
Interrogation	linguistique	Style d'écriture	booléen
Nb of premises	linguistique	Argumentation	$0 < x < n$
conclusion ?	linguistique	Argumentation	booléen
ArgumentInFirstSentence	linguistique	Argumentation	booléen
Advising	linguistique	Argumentation	$0 < x < n$
Advising	linguistique	Accord	booléen
Advising	linguistique	Désaccord	booléen

Le module de détection SOMA

Etape 2 - Représentation du contenu (préparation pour modèle d'apprentissage)

Le module de détection SOMA

Etape 2 - Représentation du contenu (préparation pour modèle d'apprentissage)

Valeur des critères extraits → Entrée de l'algorithme d'apprentissage

Pour chaque critère, on a :

- 0 si critère non détecté
- *n* sinon, où *n* = nombre de fois critère détecté dans le message

Le module de détection SOMA

Etape 2 - Représentation du contenu (préparation pour modèle d'apprentissage)

Valeur des critères extraits → Entrée de l'algorithme d'apprentissage

Pour chaque critère, on a :

- 0 si critère non détecté
- *n* sinon, où *n* = nombre de fois critère détecté dans le message

Préparation pour l'étape 3 - Application d'un modèle d'apprentissage

Génération d'une matrice de valeurs correspondant à l'ensemble de features du modèle d'apprentissage.

Le module de détection SOMA

Étape 3 - Application d'un modèle d'apprentissage

Le module de détection SOMA

Étape 3 - Application d'un modèle d'apprentissage

Hybridation de l'approche

- Tâche : classification supervisée de messages selon deux classes « influenceur », « non-influenceur »

Le module de détection SOMA

Étape 3 - Application d'un modèle d'apprentissage

Hybridation de l'approche

- Tâche : classification supervisée de messages selon deux classes « influenceur », « non-influenceur »
- Annotation manuelle d'un jeu d'entraînement

Le module de détection SOMA

Étape 3 - Application d'un modèle d'apprentissage

Hybridation de l'approche

- Tâche : classification supervisée de messages selon deux classes « influenceur », « non-influenceur »
- Annotation manuelle d'un jeu d'entraînement
- Algorithme d'apprentissage : « Random Forests » - Forêts d'arbres décisionnels (double avantage)

Le module de détection SOMA

Étape 3 - Application d'un modèle d'apprentissage

Hybridation de l'approche

- Tâche : classification supervisée de messages selon deux classes « influenceur », « non-influenceur »
- Annotation manuelle d'un jeu d'entraînement
- Algorithme d'apprentissage : « Random Forests » - Forêts d'arbres décisionnels (double avantage)
  - Prédiction de probabilité pour classification (classe définie par la majorité des arbres)

Le module de détection SOMA

Étape 3 - Application d'un modèle d'apprentissage

Hybridation de l'approche

- Tâche : classification supervisée de messages selon deux classes « influenceur », « non-influenceur »
- Annotation manuelle d'un jeu d'entraînement
- Algorithme d'apprentissage : « Random Forests » - Forêts d'arbres décisionnels (double avantage)
  - Prédiction de probabilité pour classification (classe définie par la majorité des arbres)
  - Calcul d'ordre d'importance sur les caractéristiques d'entrée (features)

Le module de détection SOMA

Étape 4 - Calcul de score d'influence par utilisateur

Approche préliminaire :

Soit

- $U = \{u_1, u_2, \ldots, u_n\}$: l'ensemble d'utilisateurs dans un réseau social,
- $S_u = \{s_1, s_2, \ldots, s_{K_u}\}$: l'ensemble de scores par message d'un utilisateur donné u,
- $K_u =$ nombre de messages postés par l'utilisateur

$$Influence(u) = \frac{\frac{1}{K_u} \sum_{i=1}^{K_u} s_i}{\max_{u'} \frac{1}{K_{u'}} \sum_{j=1}^{K_{u'}} s_j}$$

La détection d'influenceurs dans les médias sociaux

État de l'art

Notre approche

- Les caractéristiques d'influence
- Approche hybride implémentée

Expérimentation

- Les jeux de données
- Évaluation
- Résultats

Constitution des jeux de données

Un corpus de 5000 fils de discussion (forums de cosmétique)

Répartition du corpus

- 1000 fils : Annotation manuelle (message influenceur ou pas)
- 1000 fils : Développement des modules de règles linguistiques
- 3000 fils : Evaluation du modèle d'apprentissage

Constitution des jeux de données

Un corpus de 5000 fils de discussion (forums de cosmétique)

Répartition du corpus

- 1000 fils : Annotation manuelle (message influenceur ou pas)
- 1000 fils : Développement des modules de règles linguistiques
- 3000 fils : Evaluation du modèle d'apprentissage

Protocole d'annotation manuelle

- Guide d'annotation décrivant les caractéristiques d'influence
- Les annotateurs suivent le guide et indiquent « influenceur » ou « non-influenceur » pour chaque message
- Corpus sert de jeu d'entraînement pour le modèle d'apprentissage

VISEO

Evaluation : pertinence de critères linguistiques

Deux techniques adoptées :

1. Déployer deux versions du module

- Entraîner deux modèles d'apprentissage : avec et sans critères linguistiques

2. Calculer un ordre d'importance sur les critères

Evaluation : pertinence de critères linguistiques

Deux techniques adoptées :

1. Déployer deux versions du module

- Entraîner deux modèles d'apprentissage : avec et sans critères linguistiques
- Comparer les performances pour déterminer leur pertinence

2. Calculer un ordre d'importance sur les critères

Evaluation : pertinence de critères linguistiques

Deux techniques adoptées :

1. Déployer deux versions du module

- Entraîner deux modèles d'apprentissage : avec et sans critères linguistiques
- Comparer les performances pour déterminer leur pertinence

2. Calculer un ordre d'importance sur les critères

- Score d'importance généré par l'algorithme de forêts d'arbres décisionnels

Evaluation : pertinence de critères linguistiques

Deux techniques adoptées :

1. Déployer deux versions du module

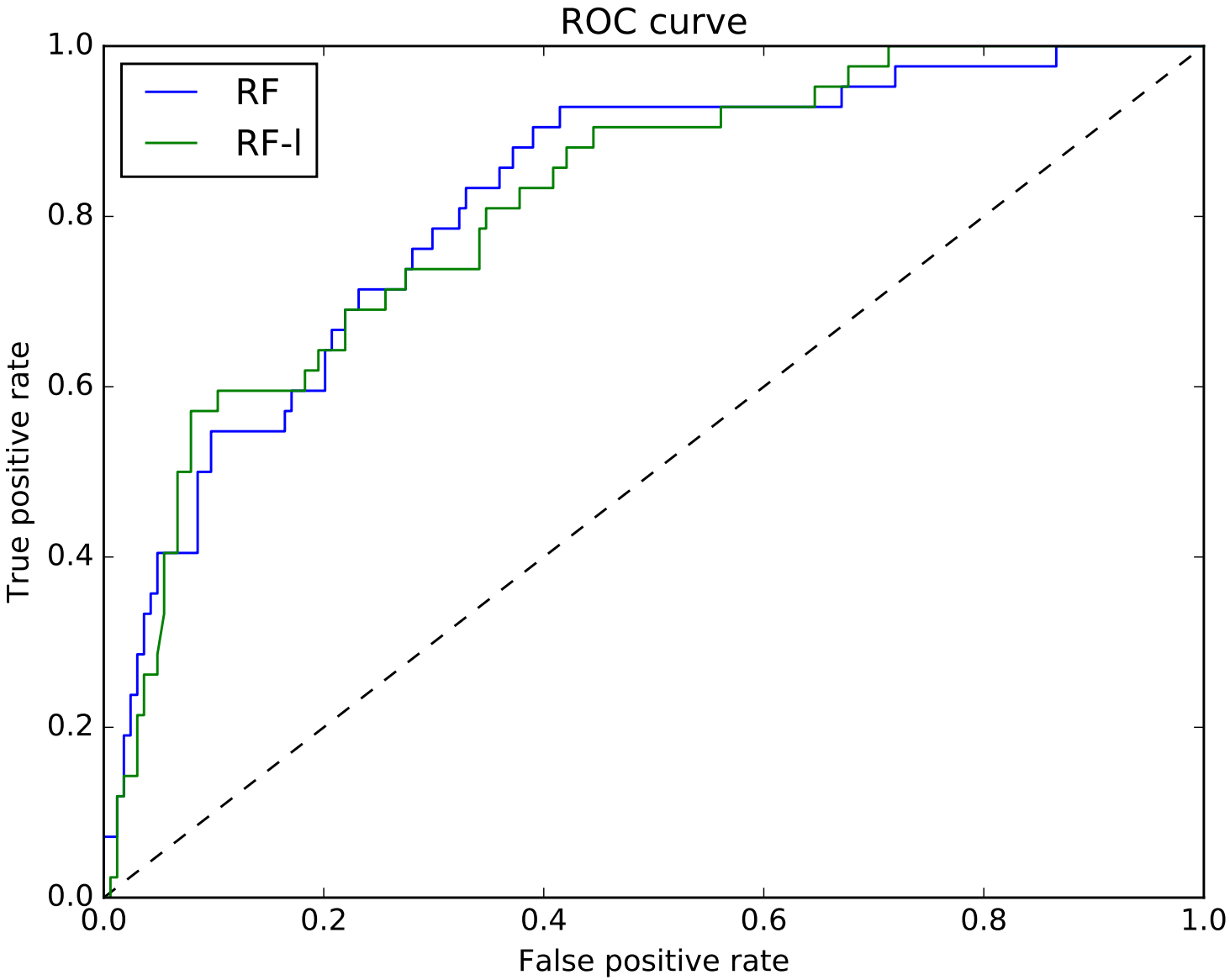
- Entraîner deux modèles d'apprentissage : avec et sans critères linguistiques
- Comparer les performances pour déterminer leur pertinence

2. Calculer un ordre d'importance sur les critères

- Score d'importance généré par l'algorithme de forêts d'arbres décisionnels
- Classer les critères par importance pour déterminer leur pertinence

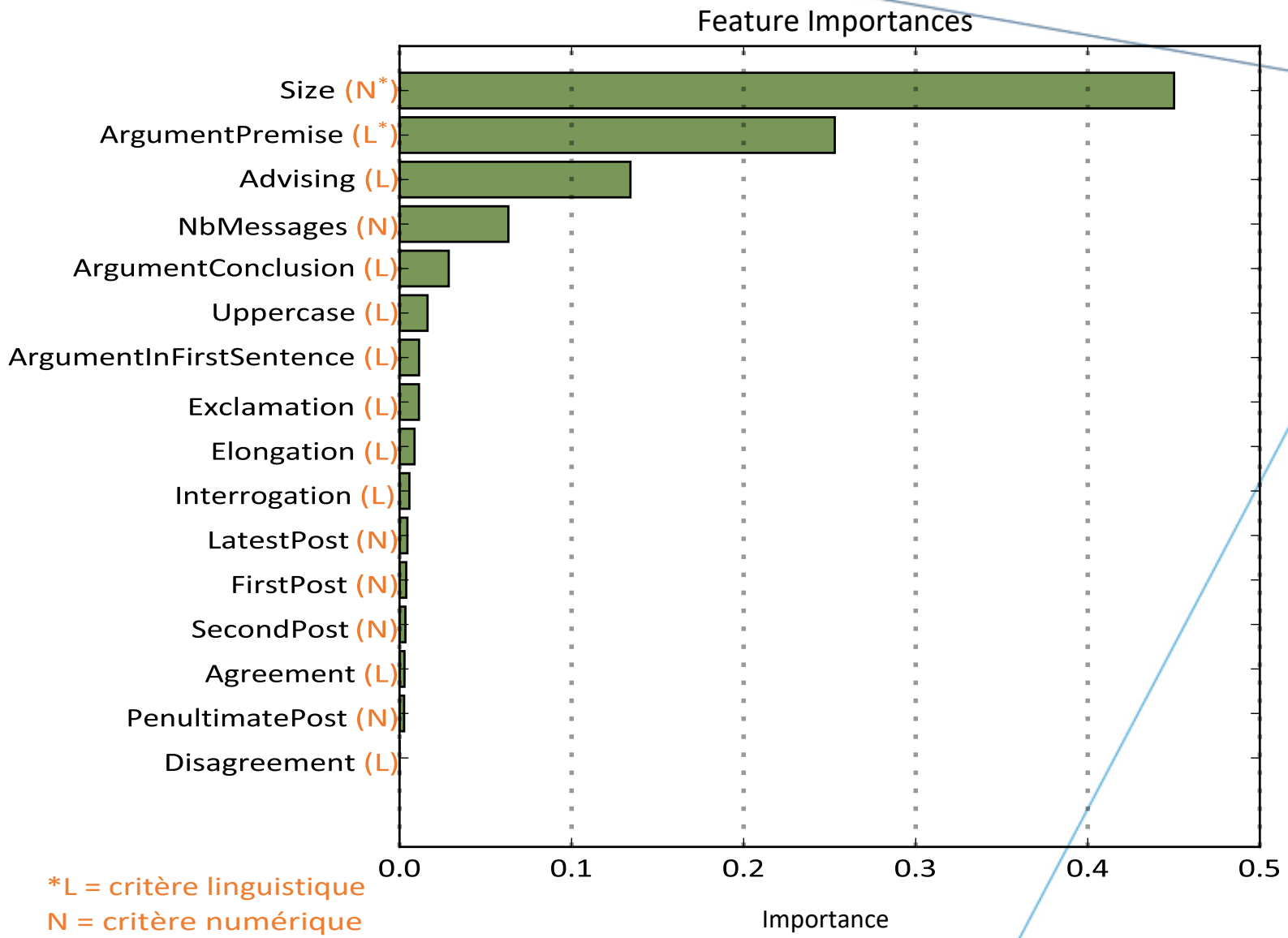
Evaluation

Courbes ROC-AUC



Evaluation

Classement de critères par importance



*L = critère linguistique
N = critère numérique

Conclusion et la suite ...

Conclusion

- Pertinence d'extraction de critères linguistiques confirmée

Conclusion et la suite ...

Conclusion

- Pertinence d'extraction de critères linguistiques confirmée
- L'hybridation d'approches symboliques et par apprentissage est mieux que l'approche par apprentissage seule

Conclusion et la suite ...

Conclusion

- Pertinence d'extraction de critères linguistiques confirmée
- L'hybridation d'approches symboliques et par apprentissage est mieux que l'approche par apprentissage seule

Conclusion et la suite ...

Conclusion

- Pertinence d'extraction de critères linguistiques confirmée
- L'hybridation d'approches symboliques et par apprentissage est mieux que l'approche par apprentissage seule

La suite

- Amélioration des modules linguistiques : extraction plus précise d'informations

Conclusion et la suite ...

Conclusion

- Pertinence d'extraction de critères linguistiques confirmée
- L'hybridation d'approches symboliques et par apprentissage est mieux que l'approche par apprentissage seule

La suite

- Amélioration des modules linguistiques : extraction plus précise d'informations
- Création d'un Gold Standard adapté : guide d'annotation qui incorpore les critères linguistiques

Conclusion et la suite ...

Conclusion

- Pertinence d'extraction de critères linguistiques confirmée
- L'hybridation d'approches symboliques et par apprentissage est mieux que l'approche par apprentissage seule

La suite

- Amélioration des modules linguistiques : extraction plus précise d'informations
- Création d'un Gold Standard adapté : guide d'annotation qui incorpore les critères linguistiques
- Analyse approfondie pour faire évoluer le calcul du score d'influence

VISEO

Conclusion et la suite ...

Conclusion

- Pertinence d'extraction de critères linguistiques confirmée
- L'hybridation d'approches symboliques et par apprentissage est mieux que l'approche par apprentissage seule

La suite

- Amélioration des modules linguistiques : extraction plus précise d'informations
- Création d'un Gold Standard adapté : guide d'annotation qui incorpore les critères linguistiques
- Analyse approfondie pour faire évoluer le calcul du score d'influence
- Améliorer la détection d'influenceurs par une analyse structurelle du réseau d'interactions : thèse démarrée en Avril'17

VISEO