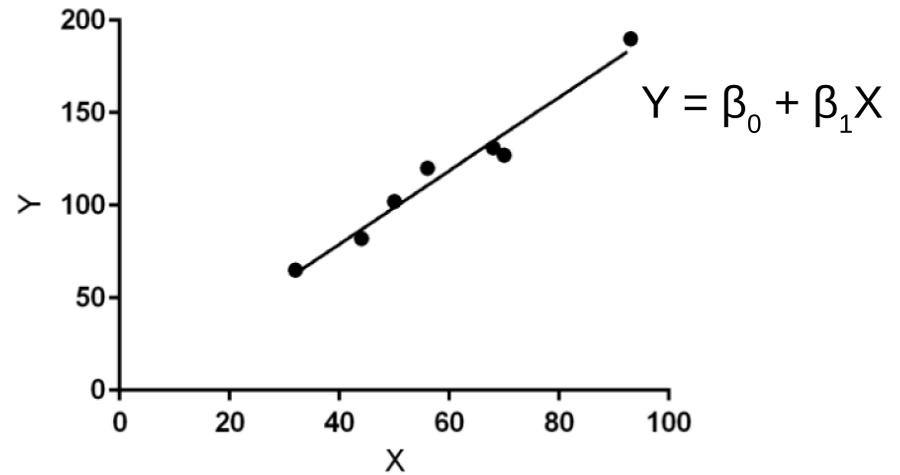


Jacobs Alexandre, Bonaert Gregory, Ruggoo Prateeba, Rouma Florian, Engelman David, Engelman Benjamin

## MODÈLE (RIDGE CLASSIFIER)

Ce type de modèle est une amélioration de la régression linéaire. En plus de minimiser les écarts entre les valeurs prédites et réelles, il force les coefficients  $\beta$  à être plus petits et donc minimise l'impact du bruit dans les problèmes avec un grand nombre de features.



## DATASET

Un dataset composé de 60.000 articles de presse, sur des sujets divers, labellisés faux ou vrai. 80 % du dataset a été utilisé pour l'entraînement du modèle et les 20 % restants pour le tester.

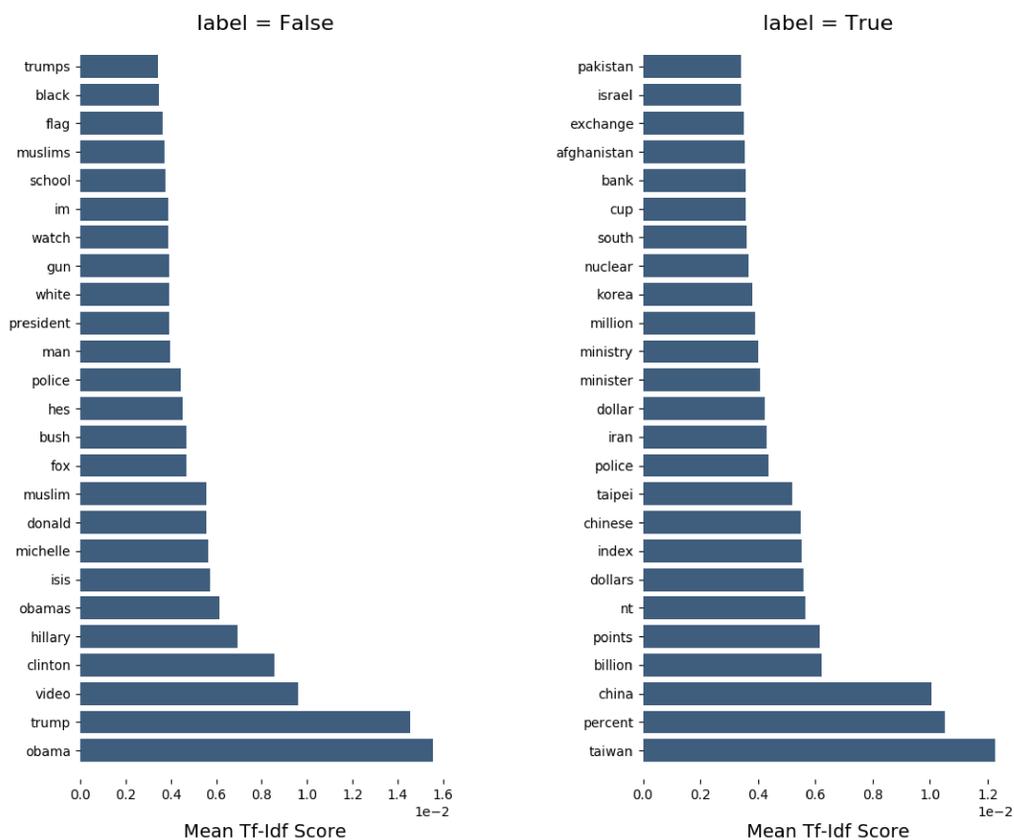
## TF-IDF

$$w_{x,y} = tf_{x,y} \times \log\left(\frac{N}{df_x}\right)$$

**TF-IDF**  
Terme X dans le document Y

$tf_{x,y}$  = Fréquence du terme X dans le document Y  
 $df_x$  = Nombre de documents contenant le terme X  
N = Nombre total de documents

### Termes les plus associés aux Fake news et aux Real News



### Meilleures combinaisons de modèles et features

	Modèle	Features	Précision
1	Ridge Classifier	Tf-Idf Ponctuation Pronoms	0.9737
2	Ridge Classifier	Tf-Idf Ponctuation Pronoms Sentiments	0.9730
3	Ridge Classifier	Tf-Idf Ponctuation Sentiments	0.972
4	Ridge Classifier	Tf-Idf Ponctuation	0.971
5	Ridge Classifier	Tf-Idf Sentiments	0.9704
6	Ridge Classifier	Tf-Idf	0.9702
6	Passive-Aggressive	Tf-Idf Sentiments Pronoms	0.9702
8	Passive-Aggressive	Tf-Idf Pronoms	0.96
9	Logistic regression	Tf-Idf Text_Count	0.95
10	Logistic regression	Tf-Idf Text_Count Sentiment	0.94