L'éxploitation des techniques de régression pour l'évaluation de la crédibilité des tweets

Hamda Slimi, Ibrahim Bounhas
Yahya Slimani

Laboratoire LISI
http://www.jarir.tn
École Nationale des Sciences de l'Informatique
Université de la Manouba, La Manouba 2010, Tunisie
hamda.slimi@ensi-uma.tn,
ibrahim.bounhas@gmail.com
yahya.slimani@gmail.com

Résumé. La crédibilité de l'information diffusée sur les réseaux sociaux est de plus en plus suspectée. En effet, accusés d'avoir servi de plate-formes de propagande pendant des événements de grande envergure (Elections américaine, BREXIT, etc.), ces réseaux ne cessent pas de perdre en crédibilité. Dans ce contexte, plusieurs travaux de recherche ont été proposés pour faire face à ces manipulations de l'information. Dans ce travail, nous présentons une approche permettant l'évaluation de la crédibilité des informations sur le réseau social Twitter. L'approche modélise la crédibilité comme une valeur continue qui prend en compte plusieurs caractéristiques des tweets, afin d'analyser et de prédire la crédibilité de ces derniers. Plusieurs variantes de notre approche ont été testées afin de valider l'apport de notre proposition. Les résultats obtenus ont montré la qualité et l'intérêt de notre approche.

1 Introduction

Twitter était considéré comme un réseau social régulier où ses utilisateurs partagent leurs activités quotidiennes. Toutefois, plusieurs facteurs ont contribué à l'évolution de Twitter en une plateforme de renommée mondiale. La nature des interactions sociales, où Twitter permet aux utilisateurs de se suivre mutuellement sans demande de permission explicite, contrairement à Facebook qui exige l'approbation d'une demande d'amitié entre utilisateurs. C'est ainsi que Twitter est devenu plus centré autour d'une structure appelée graphe d'intérêt, où les utilisateurs suivent d'autres qui leurs fournissent des informations ayant une valeur ajoutée (et qui ont donc un intérêt). De plus, certaines caractéristiques d'un tweet, tels que le hashtag (topic) et sa taille (280 caractères) permettent d'obtenir une description correcte et brève d'informations dignes d'intérêt. Dans (Kwak et al., 2010) les auteurs ont collecté et analysé 106 millions de tweets décrivant 4262 sujets de tendances. Cette analyse leur a permis de constater que 85 % des sujets ont fait la une des journaux et tout tweet retweeté à atteint jusqu'à 1000 utilisateurs, quel que soit le nombre d'abonnés du compte de son auteur originel. Ces caractéristiques

incitent à une diffusion rapide et efficace de l'information, sans être capable de fournir une information sur sa crédibilité.

La définition de la notion de crédibilité est sensible au contexte, mais l'idée communément acceptée est que la crédibilité est liée à la « croyabilité » (believability) d'une information ou d'une source(Flanagin et Metzger, 2007). Le contexte de notre étude s'intègre dans le domaine de la science de l'information ou l'évaluation de la crédibilité exige de déterminer à quel point l'information est « bonne », utile et pertinente pour aboutir à tel ou tel objectif. Les facteurs qui influent la crédibilité de l'information sont la notoriété de son auteur, du message et du média (Li et Suh, 2015). La crédibilité de l'auteur dépend de son expertise dans le domaine et de la probabilité qu'il fournit (habituellement) des informations crédibles. La crédibilité d'un message fait référence à la crédibilité perçue du message communiqué, qui dépend de la qualité et de l'exactitude de l'information. La crédibilité du média fait référence à la perception subjective de l'utilisateur du média sur sa fiabilité (Li et Suh, 2015).

2 L'état de l'art

Pour remédier à cette situation, différentes approches d'évaluation de crédibilité des tweets ont été proposées dans la littérature (Castillo et al., 2011, 2013; Kang et al., 2012). Dans cette section, nous nous intéresserons principalement sur les approches basées sur la classification et celles basées sur la propagation car elles sont étroitement liées à notre approche. En ce qui concerne les approches basées sur la classification, celles-ci considèrent la crédibilité en tant que valeur binaire : un tweet est crédible ou non crédible. De telles approches utilisent un classifieur pour apprendre un modèle basé sur un ensemble de caractéristiques relatives au contenu et à l'auteur d'un tweet.

Dans (Castillo et al., 2011), les auteurs ont collecté un ensemble de tweets à propos de 2500 topics à travers l'interface twitter streaming API. Ensuite, ils ont adopté la stratégie Best First pour sélectionner les caractéristiques du tweet et de son auteur, qui permettent une évaluation pertinente de sa crédibilité. Enfin, les auteurs ont testé une variété d'algorithmes d'apprentissage à travers des mesures de performance telles que le rappel et la précision. Les résultats ont montré que l'arbre de décision j48 donnait les meilleurs résultats avec 86% de précision, et 86% de rappel. Dans (Castillo et al., 2013), les auteurs ont extrait des tweets décrivant le tremblement de terre chilien tout en se focalisant sur la collecte de métadonnées spatiotemporelles. Ils ont évalué la performance de plusieurs algorithmes d'apprentissage tels que ceux basés sur les méthodes bayésiennes, sur la régression logistique, J48, Random Forest et ceux basés sur le méta-apprentissage. Ils ont trouvé que J48, Random Forest et le méta-apprentissage ont surpassés les autres algorithmes en atteignant une précision de 88% et un rappel de 89%. Dans (Kang et al., 2012), les auteurs ont proposé un modèle d'évaluation des utilisateurs reposant sur des indicateurs du réseau social pour calculer la crédibilité. Ainsi qu'un modèle de contenu qui adopte une approche probabiliste basée sur la langue et autres propriétés de tweets qui ont tendance à conduire à des réactions positives telles que le nombre de « Retweets » et « j'aime ». Ensuite, les auteurs ont testé la performance de chaque modèle ainsi qu'un modèle hybride qui combine à la fois le modèle social et le modèle basé sur le contenu. Les résultats ont montré que le modèle social est plus performant que les autres modèles dans la prédiction de la crédibilité avec une précision de 87% et un rappel de 88%.

Sur la base d'une revue de la littérature, nous pouvons noter que les approches basées sur la classification, SVM Rank et l'arbre de décision J48 donnent les meilleurs résultats. Ceci s'explique par le fait que la plupart des caractéristiques sont binaires, telles que la présence d'url, le sentiment d'un du tweet (négatif / positif), la présence de points d'interrogation. En outre, nous avons remarqué que les approches basées sur la propagation sont plus performantes, car elles tiennent compte l'aspect social d'un tweet et de son auteur. Cependant, les deux approches considèrent la crédibilité comme une valeur binaire : un topic est soit crédible ou non. En outre, la plupart des travaux de la littérature se concentrent sur l'évaluation de la crédibilité d'un topic (hashtag) et néglige l'évaluation (tweet). À notre connaissance, aucune approche n'a tenté de calculer la crédibilité d'un tweet en tant que valeur continue.

3 Problématique

Le contexte est un facteur majeur dans toute évaluation de la crédibilité. Par exemple, l'évaluation de la crédibilité d'un contenu web diffère de l'évolution d'un contenu hors ligne. Ce dernier dépend uniquement de l'information et de sa source, tandis que le premier dépend de multiples facteurs, à savoir le support, le format, l'autorité et l'expertise de la source. En twitter, un tweet est caractérisé par plusieurs caractéristiques. Premièrement, les caractéristiques de contenu telles que la longueur du tweet, le sentiment tweet, la présence d'URL, la présence de points d'interrogation et d'émoticon. Deuxièmement, les caractéristiques de l'utilisateur telles que son expertise, le nombre d'abonnés et l'âge de son compte. Ces caractéristiques diffèrent par leur impact. Certaines d'entre elles augmentent la crédibilité des tweets alors que d'autres la diminuent. Dans notre étude, nous avons sélectionné trois caractéristiques de contenu pour prédire la crédibilité des tweets. Nous avons opté pour une approche minimaliste qui ne nécessiterait que le contenu du tweet tout en produisant des résultats relativement précis.

4 Sélection de caractéristiques

Pour obtenir le vecteur de prédicteurs, nous devons sélectionner les caractéristiques de tweet les plus pertinentes qui caractériseraient et impacteraient la crédibilité. Nous partons de l'hypothèse que les tweets crédibles respectent les règles de la grammaire a un certain degré, ont le même sentiment que celui émis par l'événement auquel il appartienne et fournissent un lien externe (URL) pour garantir la fiabilité de l'information. Les caractéristiques que nous avons sélectionnées pour l'évaluation de crédibilité sont :

— Évaluation grammaticale : un tweet digne de mention et crédible doit respecter dans une certaine mesure les règles de grammaire afin que les informations transmises soient compréhensibles. Pour évaluer la grammaire nous avons fait recours a Grammarcheck qui est un wrapper Python pour LanguageTool.

— Accord du sentiment tweet avec le sentiment de l'événement : les tweets sont assemblés dans des topics à l'aide d'hashtag. Comme indiqué dans l'état de l'art, un sentiment de tweet crédible devrait être similaire au sentiment général des événements. Par exemple, un événement de crise aura probablement la majorité des tweets décrivant une situation d'urgence ou apportant une consolation (sentiment négatif). Par conséquent, tout tweet qui émet un sentiment positif risque de ne pas être crédible. Pour évaluer le sentiment

des tweets nous avons utilisé TextBlob. TextBlob est une bibliothèque python dédié au traitement de données textuelles.

— Présence d'un lien : la plupart des travaux existants Castillo et al. (2011, 2013) indique que la présence d'un lien externe (URL) est un signe de crédibilité du tweet, car la majorité des tweets crédibles contiennent un lien faisant référence à un média d'actualité (JSC, BBC, SKY news).

5 Expérimentation

Dans cette section, nous présentons les différents aspects de nos expérimentations et décrivons à la fois l'ensemble de données que nous avons exploité et les outils sur lesquels nous nous sommes appuyés pour évaluer les algorithmes de régression.

5.1 Description du l'ensemble de données

Nous avons utilisé CREDBANK (Mitra et Gilbert, 2015), une collection de 60 millions d'identifiants de tweet regroupés dans 1049 topic, chaque sujet étant annoté par 30 annotateurs humains sur une échelle de Likert de 5 points, allant de inexacte à exacte. Pour obtenir une annotation de crédibilité unifiée pour chaque sujet, nous avons adopté la formule proposée par (Mitra et al., 2017) selon laquelle les auteurs unifiaient les annotations de crédibilité en calculant la proportion du nombre des annotateurs qui ont annoté le topic comme exacte vis-à-vis le reste.

Nous avons collecté 50k tweet. Après avoir supprimé les tweets au sentiment neutre, nous nous sommes retrouvés avec un ensemble de données contenant 12k tweet. Dans le tableau ref DDescription, nous présentons en détail le contenu de notre ensemble de données.

Topic	Nombre de tweets	Score de crédibilité
ebola white health	433	0.77
oscar pistorius because	245	0.33
host patrick neil	3937	1.0
artist vote year	751	0.56
october ebola house	43	0.4
giants game win	6778	0.93

TAB. 1 – *Description de l'ensemble de donnée*

5.2 Évaluation des résultats

Sur la base des travaux existants, nous pouvons remarquer qu'aucun travail antérieur n'avait utilisé d'algorithmes de régression pour prédire la crédibilité des tweets. Par conséquent, nous n'avons pas pu trouver de référence (Baseline) nous permettant d'évaluer les performances de notre modèle. De telles circonstances nous ont amenés à créer un baseline naïve. Le principe d'obtention de ce baseline est comme suit : nous calculons la médiane des valeurs réelles de crédibilité et on le considère comme prédiction pour chaque tweet dans le jeu de données.

Nous avons évalué les performances des modèles de régression linéaire à l'aide de deux mesures, à savoir l'erreur quadratique moyenne et le R-carré (coefficient de détermination), afin de déterminer dans quelle mesure le modèle explique toute la variabilité des données de réponse.

Algorithmes de régression	Root mean squared error	R-square
Régression lineaire multiple	0.152	-8.07
SGD Regressor	0.151	-7.51
Baseline	0.157	-0.33
Random Forest Regressor	0.05	–

TAB. 2 – *Comparaison des algorithmes de régression*

Comme indiqué en 2, les modèles de régression linéaire et la ligne de base ont atteint un RMSE presque similaire. Cependant, nous pouvons noter que R-square a une valeur négative, ce qui montre que les modèles linéaires ont eu des performances médiocres et n'ont pas pu suivre la tendance des données. Par conséquent, nous abordons l'évaluation de la crédibilité des tweets en tant que problème non linéaire et nous nous référons au Random Forest Regressor. Après avoir évalué les performances du RFR, nous pouvons clairement constater que celui-ci a dépassé les modèles linéaires en atteignant un taux d'erreur RMSE plus petit.

Afin de sélectionner le nombre optimal d'estimateurs (T), nous calculons out of bag score (Breiman) pour différents nombres d'arbres. Nous avons remarquer que T = 800 obtient les meilleurs résultats avec un out of bag score égal à 86,46 %.

6 Conclusion

Dans ce papier, nous avons proposé une approche qui permet de déterminer le niveau de crédibilité d'un tweet, contrairement aux travaux existants qui se contentent seulement de déterminer si un tweet est crédible ou non. Notre approche se caractérise par l'exploitation de la méthode de régression pour l'évaluation de la crédibilité en tant que valeur continue. Un tel choix se justifie par notre hypothèse qui considère que si le jugement de crédibilité est de type binaire, cela aura pour conséquence d'ignorer la sémantique de la crédibilité de l'information.

Du fait que notre approche repose sur la méthode de régression, nous l'avons évalué sur 12000 tweets sur lesquels nous avons testé une panoplie d'algorithmes de régression linéaire et non linéaire afin de distinguer l'approche la plus appropriée (linéaire ou non linéaire). A la suite de l'évaluation des résultats des tests, nous avons remarqué que l'algorithme Random Forest Regression (RFR) est le plus performant. Par la suite, nous avons testé RFR avec différents paramètres, afin de déterminer le nombre d'arbres de décision qui permettrait d'avoir un bon niveau de crédibilité dans des temps raisonnables.

Comme travaux futurs, nous avons l'intention d'inclure des caractéristiques sociales dans notre approche et évaluer la crédibilité des utilisateurs. Nous nous proposons aussi de raffiner le score de crédibilité en fonction de la crédibilité de l'auteur et de son réseau social.

Références

Breiman, L. Out-of-bag estimation.

Castillo, C., M. Mendoza, et B. Poblete (2011). Information credibility on twitter. In *Proceedings of the 20th International Conference on World Wide Web*, pp. 675–684. ACM.

Castillo, C., M. Mendoza, et B. Poblete (2013). Predicting information credibility in time-sensitive social media. *Internet Research 23*(5), 560–588.

Flanagin, A. J. et M. J. Metzger (2007). The role of site features, user attributes, and information verification behaviors on the perceived credibility of web-based information. *New Media & Society 9*(2), 319–342.

Kang, B., J. O'Donovan, et T. Höllerer (2012). Modeling topic specific credibility on twitter. In *Proceedings of the 2012 ACM international conference on Intelligent User Interfaces*, pp. 179–188. ACM.

Kwak, H., C. Lee, H. Park, et S. Moon (2010). What is twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web*, pp. 591–600. AcM.

Li, R. et A. Suh (2015). Factors influencing information credibility on social media platforms: Evidence from facebook pages. *Procedia computer science 72*, 314–328.

Mitra, T. et E. Gilbert (2015). Credbank: A large-scale social media corpus with associated credibility annotations. In *ICWSM*, pp. 258–267.

Mitra, T., G. P. Wright, et E. Gilbert (2017). A parsimonious language model of social media credibility across disparate events. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, pp. 126–145. ACM.

Summary

Microblogs are growing at an overwhelming rate due to the liberty of content generation. This led to a diversity of misleading use, such as digitally manipulated content and reposting of real content in a wrong context. In order to remedy such deficiencies numerous solutions were proposed in the field of credibility evaluation of information shared on microblogs. In this paper, we review previous works about tweet credibility assessment and introduce a new approach that allows for the measurement of tweet credibility using Random Forest Regressor. In our approach, we consider credibility as a continuous value instead of a binary one and rely on three features namely; accordance between tweet and event sentiment, tweet grammar, presence of URL to predict credibility. These features were considered by the literature as suitable for credibility prediction. Based only on content features, our approach is able to predict credibility values with an out of bag score of 86.46 %.