Apprentissage déséquilibré pour la détection des signaux de l'implication durable dans les conversations en parfumerie

Yizhe WANG[1] Damien NOUVEL[2] Marguerite LEENHARDT [1] Gaël PATIN[1]
(1) XiKO, 87 rue Gabriel Péri, Paris, France
(2) ERTIM, 2 rue de Lille, Paris, France
wangyizhe0201@gmail.fr, damien.nouvel@inalco.fr,
marguerite.leenhardt@xiko.fr, gael.patin@xiko.fr

RÉSUMÉ

Une simple détection d'opinions positives ou négatives ne satisfait plus les chercheurs et les entreprises. Le monde des affaires est à la recherche d'un «aperçu des affaires». Beaucoup de méthodes peuvent être utilisées pour traiter le problème. Cependant, leurs performances, lorsque les classes ne sont pas équilibrées, peuvent être dégradées. Notre travail se concentre sur l'étude des techniques visant à traiter les données déséquilibrées en parfumerie. Cinq méthodes ont été comparées : Smote, Adasyn, Tomek links, Smote-TL et la modification du poids des classe. L'algorithme d'apprentissage choisi est le SVM et l'évaluation est réalisée par le calcul des scores de précision, de rappel et de f-mesure. Selon les résultats expérimentaux, la méthode en ajustant le poids sur des coût d'erreurs avec SVM, nous permet d'obtenir notre meilleure F-mesure.

ABSTRACT

Automatic detection of positive enduring involvement signals in fragrance products reviews

Opinion mining have been widely studied in natural language processing. Nevertheless, in recent years, a simple detection of positive or negative opinion can no longer satisfy researchers and companies. The business world is looking for "business insights". A lot of machine learning algorithms can be used to deal with the problem. However, their performance on imbalanced data can be degraded. In this paper, we focus on the study of techniques aimed at treating imbalanced data in the perfume sector. Five methods were compared : Smote, Adasyn, Tomek links, Smote-TL and changing weight of the class. The selected standard classifier is SVM and precision, recall and F-mesure scores are calculated for evaluation. According to experimental results, the method by adjusting the cost weight allows us to get our best F-measure.

MOTS-CLÉS : fouille d'opinions, classification asymétrique, SVM, ré-échantillonnage, apprentissage sensible aux coûts.

KEYWORDS: opining mining, imbalanced classification, SVM, resampling, cost sensitive learning.

1 Introduction

Aujourd'hui, les entreprises attendent plus qu'une simple détection d'opinion positive ou négative, mais des appréciations plus fines comme l'intention d'achat, la préférence pour les produits, la fidélité à la marque, y compris l'implication durable qui aide les experts en marketing à trouver l'explication du comportement de rachat au niveau individuel. Bien que la détection de ces signaux

liés au marketing ne soit pas encore beaucoup étudiée en fouille d'opinions, de nombreux algorithmes d'apprentissage automatique peuvent être utilisés pour ce faire : réseaux de neurones, arbres de décision, machines à vecteurs de support, etc.

Dans notre travail, nous considérons l'implication durable comme un signal binaire, présent ou absent. Nous cherchons à résoudre le problème sur un corpus déséquilibré dédié à la parfumerie, en comparant la performance des différentes techniques. Les machines à vecteur de support sont utilisées comme algorithme de classification. Notre hypothèse est qu'un algorithme d'apprentissage sensible aux coûts devrait surpasser les méthodes de ré-échantillonnage.

Nous présenterons d'abord le contexte général et notre positionnement. Ensuite, nous étudierons les techniques que nous avons choisies pour résoudre nos problèmes sur la classification asymétrique. La partie suivante est consacrée à la présentation de notre corpus et nos règles d'annotation. Les résultats seront montrés dans la partie suivante et nous terminerons par une discussion suivie par la conclusion.

2 Contexte général

2.1 Notion d'implication durable

L'implication est une notion issue de la psychologie qui interprète l'implication d'un individu en étudiant sa relation avec une autre personne, une cible ou un sujet. Bien qu'elle soit étudiée depuis plus de 30 ans en marketing, ce concept reste difficile à appréhender en raison de son interdépendance avec des concepts variés d'autres disciplines. Néanmoins, en marketing, il fait consensus que cette implication est une variable intrinsèque au niveau individuel, assimilée à l'attachement personnel aux objectifs ou aux événements (Abdolvand & Nikfar, 2012). Il y a trois types de classification de l'implication en marketing, dont le classement par nature qui a été proposée par Rothschild en 1975 et est acceptée assez largement : l'implication durable (EI) et l'implication situationnelle (SI).

Contrairement à la SI qui est liée à la situation temporaire du consommateur à l'égard du produit, l'EI est considérée comme un état stable du consommateur auprès d'un produit (Houston, 1978). Elle représente un niveau d'intérêt ou d'attachement d'un individu envers un produit à long terme. D'après Valette-Florence (Valette-Florence, 1989), l'EI se réfère à la fois à l'expérience ou à la connaissance antérieure du produit et aux valeurs auxquelles adhèrent l'individu. C'est-à-dire que l'EI sera positive pour une personne qui a testé ou utilisé un produit et qui a envie de continuer à l'utiliser pour longtemps ou qui lui porte une admiration intense.

2.2 Travaux précédents

Deux types de méthodes sont souvent utilisées en fouille d'opinions : l'approche symbolique basée sur le lexique et l'approche statistique en utilisant l'apprentissage automatique. L'approche basée sur le lexique utilise généralement une liste de mots ou d'expressions qui portent sur les opinions ou les sentiments des humains. On peut aussi utiliser les listes des termes contenant des sentiments qui existent déjà, comme SentiWordNet, SenticNet et HowNet. Les méthodes statistiques procèdent par classification, l'étape essentielle étant de choisir les caractéristiques, lexicales, syntaxique ou sémantiques, afin de représenter les messages.

Parmi les méthodes possibles pour sélectionner les caractéristiques, TF-IDF est largement utilisé. (Su *et al.*, 2014) ont essayé d'utiliser Word2Vec puis un classifieur SVM pour des commentaires, ils obtiennent une exactitude de plus de 90%. (Le & Zuidema, 2015) propose une méthode d'analyse de sentiments en utilisant un modèle de réseaux de neurones LSTM. (Hassan, 2017), avec une méthode similaire, montre que l'utilisation de vecteurs de mots obtenus à partir d'un modèle de réseaux de non supervisé de plongements de mots comme caractéristiques d'un système RNN-LSTM peut augmenter la performance du système de fouille d'opinions.

Le problème que nous traitons dans cet article est que le nombre d'instances dans les classes est presque toujours déséquilibré. Les travaux de (Leenhardt & Patin, 2014) et Li *et al.* (2011) se basent sur une méthode d'apprentissage semi-supervisé en modifiant la technique du sous-échantillonnage pour la classification asymétrique de sentiment. Au lieu de faire le ré-échantillonnage, Krawczyk *et al.* (2014) ont créé un ensemble efficace d'arbres de décision sensibles aux coûts pour la classification asymétrique. Une nouvelle approche pour la classification des données déséquilibrées a été proposée par Zhang *et al.* (2017), qui montre les bons résultats par minimisation du coût.

3 Méthodes

La performance des algorithmes baisse en face des données déséquilibrées. Ceci peut être expliqué par le fait que ces algorithmes cherchent à minimiser le taux d'erreur, et ignorent la différence entre les différents types de classification erronée. En particulier, ils supposent implicitement que toutes les erreurs de classification représentent le même coût lors de l'apprentissage du modèle (Ganganwar, 2012), ce qui n'est pas toujours optimal. Deux types de méthodes sont à considérer : celles qui modifient les données d'apprentissage, et celles qui adaptent les algorithmes.

3.1 Approches au niveau des données

Ces approches, basées sur le ré-échantillonnage, cherchent à augmenter la fréquence de la classe minoritaire ou à diminuer celle de la classe majoritaire. Ceci est fait afin d'obtenir approximativement le même nombre d'instances pour les classes. Nous avons testé les algorithmes présentés ci-dessous.

— **Smote** : Cette méthode de sur-échantillonnage se concentre sur la classe minoritaire, qui est augmentée en créant des exemples «synthétiques». C'est l'un des algorithmes les plus utilisés pour améliorer la performance de classifieurs appliqués sur les données déséquilibrées. L'agorithme fournit un ensemble de règles simples pour générer de nouvelles données «synthétisées». Bien que chaque nouvelle donnée synthétique soit construite à partir de ses parents (la donnée choisie et l'un de ses voisins les plus proches), la donnée générée n'est jamais un double exact de l'un de ses parents.

— **Adasyn** (Adaptive synthetic sampling) : Cet algorithme a été proposé en 2008 par He *et al.* (2008). L'idée essentielle est d'utiliser une distribution pondérée pour différents groupes de la classe minoritaire en fonction de leur niveau de difficulté d'apprentissage. Plus les données sont difficiles à apprendre, plus le nombre de données synthétiques générées va être important. L'approche améliore l'apprentissage par rapport aux distributions de données de deux façons : elle réduit le biais introduit par le déséquilibre des classes et déplace de façon adaptative la limite de classification à l'égard des exemples difficiles à apprendre.

— **Tomek links** Ces liens sont des paires de données qui sont les plus proches autour de la ligne de séparation. Cela signifie que ce sont les données qui vont être les plus problématiques pour la plupart des algorithmes de classification. En supprimant ces paires de données, la séparation entre les deux classes sera élargie, de sorte que notre l'algorithme pourra faire moins d'erreurs.
— **Smote-TL** : La méthode hybride combine les approches de sur-échantillonnage et celles de sous-échantillonnage en éliminant des données dans la classe majoritaire et ajoutant des données dans la classe minoritaire afin de rééquilibrer la distribution des classes (Santoso *et al.*, 2017). L'approche Smote-TL est la combinaison des algorithmes Smote et Tomek Links. Elle a d'abord été utilisée pour améliorer la classification des exemples sur le problème de l'annotation des protéines en bioinformatique (Batista *et al.*, 2003).

3.2 Approche algorithmique

Parmi les solutions algorithmiques (qui ne modifient pas les données) possibles, en tenant compte de l'effort pour la réalisation et de l'implémentation, nous avons choisi l'apprentissage sensible aux coûts qui tient compte des coûts associés aux exemples mal classés(Ting, 2002) selon leur proportion dans les données. Plus concrètement, cette méthode cible le problème d'apprentissage déséquilibré en utilisant des matrices de coûts qui décrivent les coûts de classification erronée. Pour une classification binaire, la matrice de coûts se concentre sur les faux positifs et les faux négatifs. Il n'y a aucun ajustement de poids associé aux vrais positifs et vrais négatifs car ils sont correctement identifiés.

Dans l'intention de trouver un meilleur poids pour notre corpus, on a essayé le poids de 1 fois à 15 fois à l'originel et on a pris finalement le poids 2, qui donne le meilleur résultat en terme de f-mesure.

4 Expérimentations

Pour nos expériences, nous avons sélectionné le SVM comme classifieur, puisque cet algorithme est réputé être plus précis sur des données modérément déséquilibrées. Les vecteurs de support sont utilisés pour la classification, ainsi de nombreux échantillons majoritaires éloignés de l'hyperplan de séparation peuvent être supprimés sans affecter la classification (Akbani *et al.*, 2004). Cependant, ce classifieur peut être sensible à un fort déséquilibre entre les classes.

Dans nos expériences, les scores rappportés en rappel, précision et f-mesure ne concernent que la classe ciblée. La sélection et l'optimisation du modèle sont faites selon la f-mesure.

4.1 Corpus et règles d'annotation

Notre corpus est constitué de commentaires en français sur le site d'avis beaute-test.com [1], spécialisé dans les produits de beauté. Nous avons extrait 20K verbatims sur le site et en avons sélectionné aléatoirement 9180 pour construire notre corpus. La longueur du commentaire varie largement : de 1 mot à 2514 mots. L'implication durable est une relation consommateur-objet stable basée sur les besoins inhérents du consommateur. Lorsque le comportement du consommateur tend à un objectif

1. http://www.beaute-test.com

à long terme (Ogbeide, 2014) ou reflète les sentiments durables à l'égard d'un produit ou d'une catégorie (Sirgy *et al.*, 2014). Ainsi, après des discussions avec les experts en marketing, nous avons considéré trois types d'expressions comme des signaux dénotant l'implication durable :

— l'expression de l'intention d'une utilisation prolongée :
« Je porte ce parfum depuis plus de 6 ans et je n'en demords toujours pas...j'adore »
— l'expression du rachat :
« J'ai toujours un flacon chez moi ! ! ! ! C'est mon parfum préféré que je rachèterai encore et encore ..... ! ! ! ! ! ! »
— l'expression d'attachement très forte au niveau de l'adoption :
« Une fragance fruitée subtile et enivrante je recommande ce parfum vous ne serez pas déçue l'essayer une fois c'est l'adopter ! ! ! »

La figure 1 montre des exemples d'expressions pour repérer les signaux demandés.

Expressions de l'intention d'une utilisation prolongée	**Expressions de rachat**	**Expression de l'adoptation**
ne change plus	acheterai encore	adopté
ai toujours un flacon	rachèterais	adoptation
reviens toujours	acheter à nouveau	adopter
c'est mon 4 eme flacon	reprendrai	
plusieurs flacons	y retourner	
encore fidèle		
3 fois que je l'achète		
ne peux m'en passer		
difficilement l'abandonner		
depuis bientôt 3 ans		
ne se quitte plus		
je ne le lâche plus		

FIGURE 1 – Exemples d'expressions contenant les signaux demandés

Des règles sont utilisées pour détecter l'EI. Les signaux positifs de EIs ne sont pas toujours explicites dans les avis publiés. D'ailleurs, toutes les phrases contenant les expressions qui ont les mêmes sens que les expressions prédéfinies par nos règles d'annotation ne comportent pas nécessairement les signaux à détecter. Par exemple, le verbatim « je l'ai utilisé pendant longtemps mais je l'aime plus » pourrait être détecté à tort comme positif. Si l'expression d'affection très forte au niveau de l'adoption est choisie comme un règle d'annotation au lieu de l'ensemble de concepts contenant le sentiment d'admiration, c'est parce qu'il faut éviter des ambiguïtés et que c'est mieux d'avoir une frontière plus claire avec d'autres notions en marketing, comme la préférence. De ce fait, les avis de produits avec un sentiment positif assez fort comme « Je suis tomber amoureuse de ce parfum pour les filles qui aime les notes sucrés je conseil vivement » et « Un de mes parfums préférés ! ! ! » sont annotés en

négatifs. Par ailleurs, tous les échantillons ont été annotés par une personne.

Avant d'entrer dans l'apprentissage, la tokenization et la suppression des mots vides ont été faites à l'aide de la librairie NLTK. Parmi toutes les méthodes d'extraction des caractéristiques implémentées, nous avons choisi le modèle TF-IDF.

4.2 Méthode d'évaluation

Pour faire face au problème généré par le déséquilibre entre les classes et obtenir un modèle de classification asymétrique optimal, la sensibilité et la spécificité sont habituellement adoptées pour surveiller respectivement la performance de classification sur deux classes. Cependant, parfois, on s'intéresse à la capacité de détecter efficacement une seule classe. Pour de tels problèmes, une autre paire de mesures, la précision et le rappel, est souvent adoptée. La f-mesure est aussi souvent utilisée pour intégrer la précision et le rappel dans une seule mesure pour la commodité de l'évaluation.

Notez que tous nos scores (rappel, précision et f-mesure) sont seulement ceux de la classe ciblée.

Dans nos expériences, la sélection et l'optimisation du modèle sont faites en condition de la f-mesure. Cependant, il faut savoir que dans l'application métier, c'est souvent l'objective métier ou la demande du client qui décide du choix de modèle. Par exemple, dans le travail de filtration du spam (la classe positive est spam), la sélection et l'optimisation du modèle sont souvent faites selon la précision. Parce que les faux négatifs (le spam va dans la boîte de réception) sont plus acceptables que les faux positifs.

4.3 Résultats

Le tableau 4.3 montre le changement du nombre de données après avoir utilisé les quatres techniques aux niveaux des données :

	Originel	Adasyn	TL	Smote	Smote+TL
nombre total d'échantillons	9180	13231	7243	13118	13114
classe minoritaire	907	6672	727	6559	6557

TABLE 1 – Changement du nombre d'échantillons

Le tableau 4.3 présente les résultats obtenus par utilisation du SVM sur les données et selon les différentes méthodes de prise en compte du déséquilibre des classes. Une optimisation supplémentaire du SVM a été faite en faisant varier les hyper-paramètres selon une méthode aléatoire, comme proposé par Bergstra & Bengio (2012).

4.4 Discussion

Nous voyons dans les résultats que Tomek Links fonctionne assez bien sur notre corpus. Cependant, en utilisant une méthode de sous-échantillonnage, des informations importantes risquent d'être éliminées en même temps. Il y a deux types de classification asymétrique : le déséquilibre relatif et la rareté absolue (Weiss, 2004). Notre classe minoritaire contient 907 exemples c'est un cas de déséquilibre

	Précision	Rappel	F-mesure
SVM	59.75%	62.50%	61.09%
SVM+opt	58.56%	67.76%	63%
SVM+Smote	50.41%	81.85%	62.31%
SVM+S-TL	53.42%	76.79%	63.07%
SVM+Adasyn	39.02%	**84.21%**	53.33%
SVM+TL	**66.67%**	60.53%	63.45%
SVM+couts	63.20%	67.76%	65.40%
SVM+couts+opt	60.20%	77.63%	**67.82%**

TABLE 2 – Tableau de résultats

relatif. De ce fait, après avoir supprimé des données de la classe majoritaire et avoir des classes assez équilibrées, la performance du SVM serait aussi améliorée.

En comparant l'efficacité de Tomek Links avec celui ajustant le poids de la classe minoritaire, on conclut que le dernier est plus performant sur notre corpus, puisqu'il nous permet d'obtenir notre meilleure f-mesure. C'est un désavantage connu de l'utilisation des méthodes de ré-échantillonnage L'inconvénient du sous-échantillonnage est qu'il provoque l'élimination des données potentiellement utiles. Et l'inconvénient essentiel du sur-échantillonnage est qu'en faisant des copies des exemples existants, un phénomène de sur-apprentissage risque de survenir (Weiss *et al.*, 2007). Un deuxième inconvénient du sur-échantillonnage est qu'il augmente le nombre d'exemples d'apprentissage, augmentant ainsi le temps d'apprentissage, ce qui est visible aussi dans notre expérience.

Pour nos expériences, l'apprentissage par méthode sensible aux coût est celle qui fournit les meilleurs résultats. Cependant, il est difficile de généraliser. L'expérience de Weiss *et al.* (2007) montre que l'algorithme d'apprentissage sensible aux coûts surpasse systématiquement les méthodes de ré-échantillonnage quand on se concentre exclusivement sur des ensembles de données comportant plus de 10K exemples, tandis que la technique de sur-échantillonnage semble être la meilleure méthode pour les petits ensembles de données.

Pour des travaux futurs, nous envisageons de mettre en place d'autres méthodes qui donnent de bons résultats pour la classification des classes déséquilibrées, comme la méthode *Random over-sampling*, une méthode simple mais compétitive par rapport aux autres techniques de sur-échantillonnage plus complexes (Batista *et al.*, 2004).

5 Conclusion

Dans cet article, nous avons évalué 5 algorithmes souvent utilisés en classification asymétrique afin de détecter les signaux positifs de l'implication durable dans les avis des consommateurs en parfumerie. L'algorithme Tomek Links donne les meilleurs résultats en précision, Adasyn pour le rappel. Globalement, l'approche en utilisant l'algorithme sensible aux coûts est la meilleure méthode avec une F-mesure de 67.82%. Si les méthodes de sur-échantillonnage ou de sous-échantillonnage peuvent être intéressantes pour des cas particuliers, l'approche algorithmique est celle qui apporte les meilleurs résultats.

Références

ABDOLVAND M. & NIKFAR F. (2012). Investigation of the relationship between product involvement and brand commitment.

AKBANI R., KWEK S. & JAPKOWICZ N. (2004). Applying support vector machines to imbalanced datasets. *Machine learning : ECML 2004*, p. 39–50.

BATISTA G. E., BAZZAN A. L. & MONARD M. C. (2003). Balancing training data for automated annotation of keywords : a case study. In *WOB*, p. 10–18.

BATISTA G. E., PRATI R. C. & MONARD M. C. (2004). A study of the behavior of several methods for balancing machine learning training data. *ACM Sigkdd Explorations Newsletter*, **6**(1), 20–29.

BERGSTRA J. & BENGIO Y. (2012). Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, **13**(Feb), 281–305.

GANGANWAR V. (2012). An overview of classification algorithms for imbalanced datasets. *International Journal of Emerging Technology and Advanced Engineering*, **2**(4), 42–47.

HASSAN A. (2017). Sentiment analysis with recurrent neural network and unsupervised neural language model.

HE H., BAI Y., GARCIA E. A. & LI S. (2008). Adasyn : Adaptive synthetic sampling approach for imbalanced learning. In *Neural Networks, 2008. IJCNN 2008.(IEEE World Congress on Computational Intelligence). IEEE International Joint Conference on*, p. 1322–1328 : IEEE.

HOUSTON M. J. (1978). Conceptual and methodological perspectives on involvement. *Research frontiers in marketing : Dialogues and directions*, p. 184–187.

KRAWCZYK B., WOŹNIAK M. & SCHAEFER G. (2014). Cost-sensitive decision tree ensembles for effective imbalanced classification. *Applied Soft Computing*, **14**, 554–562.

LE P. & ZUIDEMA W. (2015). Compositional distributional semantics with long short term memory. *arXiv preprint arXiv :1503.02510*.

LEENHARDT M. & PATIN G. (2014). Détecter les intentions d'achat dans les forums de discussion du domaine automobile : une approche robuste à l?épreuve des expressions linguistiques peu répandues.

LI S., WANG Z., ZHOU G. & LEE S. Y. M. (2011). Semi-supervised learning for imbalanced sentiment classification. In *IJCAI proceedings-international joint conference on artificial intelligence*, volume 22, p. 1826.

OGBEIDE O. A. (2014). Knowing your customers to serve them better : Enduring involvement approach. *Global Research Journal of Business Management*, **2**(2), 5–14.

SANTOSO B., WIJAYANTO H., NOTODIPUTRO K. & SARTONO B. (2017). Synthetic over sampling methods for handling class imbalanced problems : A review. In *IOP Conference Series : Earth and Environmental Science*, volume 58, p. 012031 : IOP Publishing.

SIRGY J., RAHTZ D. & DIAS L. (2014). Consumer behavior today. *Irvington, NY : Flatworld Knowledge Publishers*.

SU Z., XU H., ZHANG D. & XU Y. (2014). Chinese sentiment classification using a neural network tool ?word2vec. In *Multisensor Fusion and Information Integration for Intelligent Systems (MFI), 2014 International Conference on*, p. 1–6 : IEEE.

TING K. M. (2002). An instance-weighting method to induce cost-sensitive trees. *IEEE Transactions on Knowledge and Data Engineering*, **14**(3), 659–665.

VALETTE-FLORENCE P. (1989). Conceptualisation et mesure de l'implication. *Recherche et Applications en Marketing (French Edition)*, **4**(1), 57–78.

WEISS G. M. (2004). Mining with rarity : a unifying framework. *ACM Sigkdd Explorations Newsletter*, **6**(1), 7–19.

WEISS G. M., MCCARTHY K. & ZABAR B. (2007). Cost-sensitive learning vs. sampling : Which is best for handling unbalanced classes with unequal error costs ? *DMIN*, **7**, 35–41.

ZHANG C., WANG G., ZHOU Y. & JIANG J. (2017). A new approach for imbalanced data classification based on minimize loss learning. In *Data Science in Cyberspace (DSC), 2017 IEEE Second International Conference on*, p. 82–87 : IEEE.