



HAL
open science

Apprentissage automatique pour la surveillance de marques

Julie Tytgat, Guillaume Wisniewski, Adrien Bétrancourt

► **To cite this version:**

Julie Tytgat, Guillaume Wisniewski, Adrien Bétrancourt. Apprentissage automatique pour la surveillance de marques. Conférence Nationale sur les Applications Pratiques de l'Intelligence Artificielle, Jun 2021, Saint-Etienne, France. hal-03911956

HAL Id: hal-03911956

<https://hal.science/hal-03911956v1>

Submitted on 23 Dec 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Apprentissage automatique pour la surveillance de marques

J. Tytgat^{1,2}, G. Wisniewski¹, A. Bétrancourt²

¹ Université Paris Cité, LLF, CNRS 75 013 Paris, France

² IPSIDE, 31 100 Toulouse, France

julie.tytgat@etu.u-paris.fr ; guillaume.wisniewski@u-paris.fr ; a.betrancourt@ipside.com

Résumé

Les marques sont un des piliers de la propriété intellectuelle, dont la protection et la surveillance sont des véritables enjeux industriels. L'augmentation massive des dépôts de marques motive l'utilisation de techniques d'apprentissage automatique pour assister dans cette surveillance, traditionnellement assurée par des processus manuels particulièrement coûteux. Cet article présente plus formellement la tâche induite – en insistant sur les problématiques et contraintes industrielles –, puis introduit et discute les méthodes déployées et premiers résultats obtenus.

Mots-clés

Propriété intellectuelle, marques, legaltech, mesure de similarité

Abstract

Trademarks are a cornerstone of intellectual property. Their protection and monitoring are major industrial challenges. The massive increase of trademark filings advocates for the use of NLP techniques in this protection, usually performed via manual and laborious processes. This article presents more formally the resulting task – insisting on the industrial constraints –, then introduces and discusses the methods used and obtained preliminary results.

Keywords

Intellectual property, legaltech, similarity metrics

1 Introduction

L'objectif de ce travail est de présenter une nouvelle application de l'apprentissage statistique : la surveillance automatique de marques, une tâche essentielle pour la protection de la propriété intellectuelle d'une entreprise, chronophage et coûteuse.

Une marque, au sens de la propriété intellectuelle, est un « signe » qui permet de différencier les produits et services d'une entreprise de ceux de ses concurrents. Une marque protégée est un actif qui peut être exploité, vendu, transféré. C'est un actif stratégique qui synthétise l'image d'un produit ou d'un service auprès du public et participe à la valeur d'une entreprise.

Protéger une marque permet à une entreprise d'agir contre des contrefacteurs, de créer de la valeur et d'accroître la crédibilité de celle-ci face à ses partenaires. À l'inverse, un

conflit entre marques peut avoir diverses conséquences que ce soit sur l'image de la marque, ou une perte de revenus. C'est pourquoi le travail de surveillance autour de cet actif est essentiel : il faut pouvoir détecter parmi les nouveaux dépôts ceux qui risquent de remettre en question l'avantage conféré par une marque.

Dans le monde, il y a eu 13,4 millions de demandes de dépôts de marques en 2020. Cela représente 1,9 million de plus qu'en 2019 (soit 16,5% d'augmentation) et ce dans un contexte de pandémie [9]. En France, au sein de l'Institut National de la Propriété Industrielle (INPI), on compte plus de 100 000 dépôts sur la même année, soit 7,2% de plus que l'année précédente ¹.

Pour apporter des éléments de réponse à cette problématique industrielle de la surveillance, nous présentons dans cet article une tâche de classification binaire visant à détecter la similarité entre deux marques dites *verbales* (cf. 3.1), dans un cadre de surveillance. Nous commençons par présenter dans la partie 2 des travaux similaires récents. Puis un récapitulatif des critères constituant la notion de similarité est proposé à la section 3. Nous décrivons également dans cette section, une première approche visant à déterminer automatiquement cette similarité en nous inspirant de travaux antérieurs. La partie 4 explique comment ces représentations sont combinées au sein de modèles d'apprentissage. Nous présentons également en section 5 nos premiers résultats avec une analyse critique des représentations et des modèles étudiés, et identifions les problématiques scientifiques soulevées par le passage à l'échelle de ce type de méthodes dans un contexte industriel. Enfin, nous concluons dans la section 6.

2 Travaux antérieurs

L'utilisation de technologies d'intelligence artificielle dans le domaine de la propriété intellectuelle est un champ de recherche très actif [2]. Dans le cadre de la similarité entre marques en particulier, on peut identifier différentes approches se focalisant sur les multiples caractéristiques concourant à la définition d'une marque.

Concernant l'aspect sémantique, on peut citer des approches basées sur les représentations vectorielles [8] avec *word2vec* ou [1] sur ontologie. Sur le versant phonétique,

1. <https://www.inpi.fr/fr/nationales/chiffres-cles-de-la-proprietee-industrielle>

[3] utilise une approche neuronale, tandis que [4] se base sur un *soundex pondéré*.

On retrouve également des approches pluri-factorielles, par exemple [6] avec de la *fuzzy logic*, ou [8], qui combine l'analyse de la partie verbale de marques figuratives à une analyse visuelle du logo en général.

À la différence de ces précédentes contributions, nous changeons le paradigme avec lequel la notion de similarité est considérée. Nous présentons nos méthodes dans un cadre industriel de surveillance des marques, avec la volonté de passer à l'échelle, et non dans la reproduction de jugements obtenus en cours de justice, plus courants dans la littérature. Si des méthodes d'apprentissage statistique et/ou neuronale sont par ailleurs déjà utilisées pour calculer les *features* ou similarités sur un aspect de la marque, nous sommes les premiers à notre connaissance à utiliser ces méthodes pour les combiner.

3 Marques et représentation

3.1 Notion de marque

Nous allons commencer par expliquer ce qui définit une marque protégée et le risque de confusion, en nous concentrant sur les marques dites *verbales*, qui sont l'objet du travail que nous décrivons.

Des marques de différente nature peuvent être déposées : du logo à l'hologramme en passant par une marque sonore, une couleur... Ne peut être déposé ce qui ne peut faire l'objet d'une représentation graphique (comme une odeur), quelque chose désignant directement le produit ou service, ou des termes pouvant tromper le consommateur.

Une marque verbale est une combinaison de caractères typographiques standards. Cela peut être un nom, un ou plusieurs mots (Ex. : *Yoplait, Guy Degrenne*), un slogan (Ex. : *Parce que vous le valez bien (L'Oréal)*), des chiffres et/ou des lettres (Ex. : *307 (Peugeot), 24 Faubourg (Hermès)*).²

Une marque est enregistrée comme représentant des biens et/ou services spécifiques, dont la liste est l'un des principaux critères la définissant. La protection juridique à laquelle cette marque prétend s'étend donc précisément sur ces produits et services. Pour des questions de clarté et d'efficacité, cette liste est associée à un système de classification, dit de Nice, institué par l'*Arrangement de Nice concernant la classification internationale des produits et des services aux fins de l'enregistrement des marques* en 1957. Le système, mis à jour tous les 5 ans, recoupe actuellement 45 catégories, pour 34 classes de biens et 11 classes de services.

Tous ces éléments permettent de protéger un bien/service du risque de confusion correspondant à une situation dans laquelle une personne tierce utilise une marque qui ne serait pas suffisamment distincte d'une autre. Ce risque est évalué par un juge. Celui-ci prend en compte la similitude des signes, l'identité des produits et la notoriété de la marque. L'impression d'ensemble produite par ses critères auprès

d'un consommateur moyennement attentif est le niveau à partir duquel on va juger ce risque de confusion. Le domaine d'activité, symbolisé par les classes de Nice, est donc l'un des facteurs clefs de ce risque de confusion, en plus de la similitude des signes.

Réussir à déterminer la similarité entre deux marques uniquement à partir du nom de celles-ci est donc une approximation de la tâche effectuée par le juge quand il estime le risque de confusion. Lorsqu'il évalue la similarité de deux marques, notre système ne prend pas en compte les critères extrinsèques que peuvent être la notoriété de la marque, la nature du public exposé à la marque, etc. C'est pourquoi notre travail se situe toujours en amont et en complément de celui, plus complexe, fourni par des experts du domaine. Pour une marque verbale, la similitude entre les signes va s'apprécier au regard des différents critères les caractérisant. Plus précisément, il va s'agir d'une combinaison à divers degrés des aspects phonétique, graphique et sémantique.

Il faut également prendre en compte le caractère distinctif plus ou moins élevé de la marque antérieure en fonction des produits pour la désignation desquels elle a été enregistrée. À titre d'exemple, pour une marque fictive *Michel menuiserie*, enregistrée pour les classes de produits 19 et 20³, il convient de se concentrer sur la première partie du signe étant usuellement celle qui porte le caractère distinctif de la marque. Certains mots, par ex. des adjectifs mélioratifs comme *super, top*, ne constituent des éléments distinctifs et ce, quels que soient les produits et services. Le caractère distinctif doit donc s'apprécier sur la globalité du signe.

3.2 Représentation d'une marque

Nous allons maintenant présenter rapidement les différentes caractéristiques pouvant, intuitivement être utiles pour déterminer automatiquement la similarité entre deux marques. La caractéristique la plus naturelle sont les classes de Nice qui correspondent à une sur-approximation des produits et services de la marque et peuvent facilement être symbolisées sous forme de représentations vectorielles, auxquelles nous avons adjoint une fonction de similarité (cosinus, distance de Tanimoto [7]). Nous avons toutefois fait le choix de retirer cette caractéristique de notre modèle car nos premières expériences ont montré que celle-ci entraînait un sur-apprentissage. Cette approche présente également le désavantage de ne pas prendre en compte l'hétérogénéité des écarts conceptuels entre paires de classes.

Concernant le signe des marques, on se repose sur des méthodes de comparaison éprouvées : des distances d'édition (Levenshtein, Levensthein pondéré, Jaro-Winkler) et des similarités sur les chaînes de caractères en commun. Concernant ces dernières, on distinguera cependant deux approches : les mesures identifiant la plus longue sous-chaîne commune et celles comptabilisant les sous-mots communs.

2. exemples tirés de l'INPI : <https://www.inpi.fr/fr/comprendre-la-propriete-intellectuelle/la-marque/les-differents-types-de-marque>

3. Classe 19 : Matériaux de construction non métalliques, Classe 20 : Meubles, glaces (miroirs), cadres,

On choisit de ne pas développer de traitements spécifiques pour les marques de différentes tailles malgré leurs différences de fond. Une première motivation derrière ce choix est, comme pour les classes de Nice, l'observation d'une tendance à la sur-considération de cette caractéristique lors des premiers tests d'apprentissage. La seconde est que la longueur est de fait implicitement prise en compte par certaines méthodes de comparaisons du signe évoquée précédemment, comme les distances d'édition.

Pour améliorer ces mesures de similarités, il est nécessaire, même si l'on ne se trouve pas dans le cadre de marques complexes, mêlant figuratif et verbal, de prendre en compte l'aspect visuel de la marque. Typiquement, on peut citer l'utilisation d'un caractère visuellement proche pour en remplacer un autre, pour un effet de style; sans oublier qu'une ponctuation ou un nombre peut être typographié en toutes lettres. Pour gérer ces cas, nous générons lors d'une phase de pré-traitement des normalisations possibles, telles que :

- Un "!" peut être réécrit en "i"
- Un "3" peut être réécrit en "e"
- Un "4" peut être réécrit en "A", "for" ou "four"
- Un "2" peut être réécrit en "to", "two"
- Un "@" peut être réécrit en "a" ou "at"
- Un "&" peut être réécrit en "et", "and" ou "N"
- Un "+" peut être réécrit en "t" ou "plus"

Ces règles sont appliquées indépendamment de façon à générer le maximum d'écritures possibles. On sélectionne la variante maximisant le résultat de similarité lors de nos autres traitements.

De même, pour la phonétique, nous nous sommes basés sur des algorithmes déjà étudiés dans ce contexte, soundex et metaphone. On remarque néanmoins que dans le cas de marques déposées en France, on peut concevoir une prononciation française autant qu'anglaise.

4 La tâche de surveillance de marque

Pour reproduire une situation de surveillance, nous avons considéré deux sources de données présentées ci-dessous. Nous parlerons ensuite des modèles d'apprentissage que nous avons testés.

En premier lieu, nous avons accès à une base de marques issues d'un contexte industriel de surveillance. Celle-ci comporte 68 518 paires de marques identifiées comme étant similaires. Issus d'un système expert, il s'agit des exemples positifs servant de base à notre apprentissage.

La seconde source de données est la base de marques déposées fournies par l'INPI. L'institut publie hebdomadairement une mise-à-jour de cette base sous une licence libre⁴. Afin de générer des exemples négatifs nous avons, pour chaque occurrence d'une marque surveillée dans notre liste de positifs, sélectionné aléatoirement 10 marques au sein de cette base. Notre corpus contient donc au total 753 698 paires dont environ 9% sont similaires.

Cette méthodologie cherche à simuler la notion de surveillance en contexte industriel, c.-à-d. détecter d'éventuels conflits entre les marques surveillées et une liste de nouvelles marques déposées.

Notre objectif est double : nous souhaitons, dans un premier temps, être capable d'inférer à partir de cet échantillon un système capable de détecter si une paire de marque non encore observée est potentiellement en conflit; nous souhaitons également, à plus long terme, d'être capable de hiérarchiser la qualité (c.-à-d. la similarité) de ces exemples.

Ce corpus a été généré de manière à reproduire certaines des caractéristiques de la surveillance que nous avons observées : elle permet notamment de simuler le déséquilibre entre positifs (paires de marques similaires) et négatifs (paires de marques non similaires); elle a également l'avantage de reproduire l'hétérogénéité du nombre d'exemples qu'une marque en surveillance va produire : une marque très connue, ou avec un signe simple (court et/ou peu distinctif) va générer beaucoup plus de signalements qu'ils soient mérités ou erronés.

L'une d'une caractéristique de cette tâche est que le rapport de classe est déséquilibré : il y a énormément de marques déposées, pour très peu que l'on puisse considérer comme similaires. Cependant, dans ce contexte industriel, le moindre faux négatif peut s'avérer extrêmement critique. Le ratio de marques similaires ne va pas forcément être homogène : une marque verbale de seulement 3 lettres par exemple va avoir beaucoup de marques similaires sur son signe; il convient alors d'effectuer la distinction sur les produits et services.

Nous avons d'abord considéré un modèle de régression logistique, qui en plus d'être simple à mettre en place pour une première approche, a le mérite de permettre de contrôler le compromis entre rappel et précision relativement facilement (en faisant varier le seuil de décision), facteur clef dans notre tâche. Nous avons également testé un classifieur par forêt aléatoire qui, si il peut avoir tendance à sur-apprendre, présente l'avantage de fournir un certain degré d'explicabilité dans ses décisions, permettant d'avoir un premier retour critique sur les caractéristiques utilisées.

Nous avons utilisé, dans toutes nos expériences, l'implémentation de ces deux classifieurs de la bibliothèque `sklearn` et estimé les paramètres et les hyper-paramètres de manière standard.

5 Évaluation et discussion

Cette section présente nos premiers résultats et propose une discussion sur les différents points d'intérêt qu'ils soulèvent, à partir desquels sont émises des hypothèses pouvant les expliquer.

Pour évaluer les différentes méthodes évoquées, nous nous sommes servis du score F_1 et de la courbe ROC, deux métriques adaptées pour l'évaluation de classifieurs dans le contexte de problèmes déséquilibrés. Nous avons également considéré l'aspect qualitatif des résultats, toujours dans le cadre d'une coopération avec les experts.

4. <https://www.inpi.fr/fr/open-data-marques-francaises>

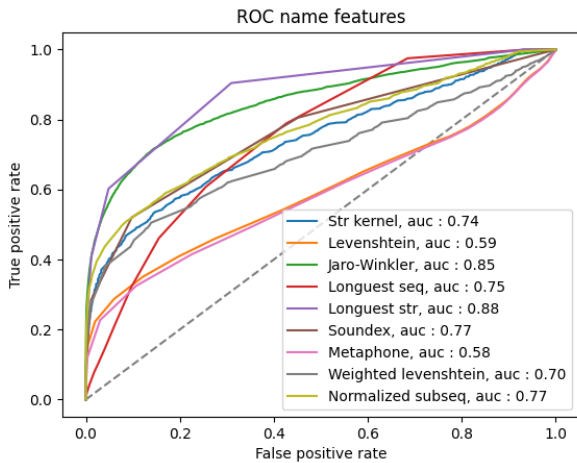


FIGURE 1 – Résultats des différentes *features* en isolation

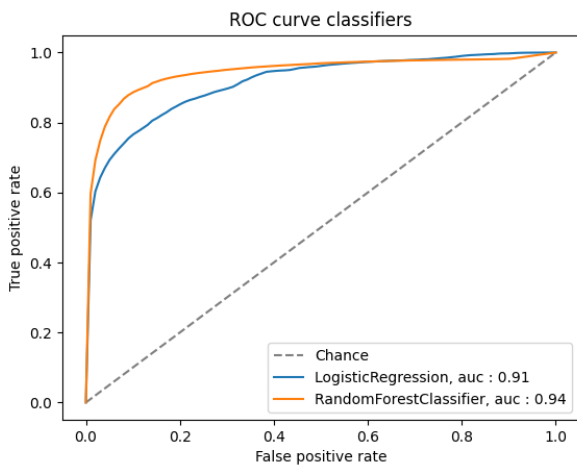


FIGURE 2 – Comparaison des modèles d'apprentissage

5.1 Features entre elles

La Figure 1 introduit, sous forme de courbes ROC, les résultats obtenus par les différentes métriques en isolation. Nous présentons et commentons ici les leçons qui peuvent en être tirées.

Distance d'édition Parmi les distances d'édition, celle obtenant les meilleurs résultats est Jaro-Winkler, loin devant la distance de Levenshtein. A noter que cette dernière s'améliore une fois pondérée⁵ : ce qui est cohérent avec les résultats de Jaro-Winkler, qui met l'accent sur le fait d'avoir des préfixes similaires. Ces résultats correspondent à ceux observés dans la littérature [4, 5].

Similarité de sous-chaîne Si on compare la présence du plus long sous-mot avec la plus longue sous-séquence, on peut penser que cette dernière est plus pertinente. Toutefois, dans une optique de minimisation des faux négatifs, la mesure du plus long sous-mot devient aussi valable. Le fait de

5. On augmente le poids des premiers mots, et le poids des mots se trouvant dans des positions proches.

normaliser la plus longue sous séquence commune par la taille de la plus longue des deux chaînes permet de minimiser le nombre de faux positifs, mais s'avère à l'inverse moins performante quand il s'agit de minimiser le nombre de faux négatifs. Les résultats d'une implémentation naïve d'une fonction de *string kernel* suivent d'ailleurs ceux de la version normalisée de la plus longue sous-séquence commune, en légèrement moins bons.

Phonétique On peut observer que soundex obtient de meilleurs résultats que metaphone. La courbe de résultats de metaphone est très proche de celle obtenue pour la distance de Levenshtein. Cela peut s'expliquer par le fonctionnement de metaphone qui va traduire les séquences observées avec des modifications plus mineures et fines que soundex. L'application de Levenshtein pour les comparer reproduit, au final, le comportement de Levenshtein sans soundex. D'autre part, soundex a initialement été conçu pour des questions de recensement, augmentant l'importance donnée au début du signe, ce qui tend à améliorer les résultats comme observé sur les différentes distances d'édition.

En tentant d'explicitier les différentes *features* pour comprendre le fonctionnement de nos classificateurs, on ne peut que constater que certaines d'entre elles, pourtant classiquement utilisées dans la littérature, n'obtiennent pas les résultats escomptés (Levenshtein). D'autres, bien que très simples, obtiennent de bons résultats (plus longue séquence commune). Bien qu'à interpréter précautionneusement, ces retours sont cruciaux pour l'explicabilité du système, et donc son acceptation par les experts amenés à l'utiliser et les retours critiques qu'ils pourraient formuler.

A partir de ces différentes composantes, on peut transformer une paire de marques en un vecteur de caractéristiques dont chaque entrée représente une similarité ou une distance précédemment mentionnée. Leur analyse nous permettent de déterminer une première approche sur la façon de constituer les vecteurs de caractéristiques. Nos systèmes d'apprentissage reçoivent comme représentation d'une paire de marque un vecteur de taille 4 (édition, sous-chaîne1, sous-chaîne2, phonétique) de manière à varier les types d'information tout en restant explicable.

5.2 Modèle d'apprentissage

La régression logistique a obtenu un score F_1 de 0.60 et la forêt aléatoire de 0.72. Les résultats sont détaillés à la Figure 2.

Le déséquilibre entre les classes est considéré à l'apprentissage par une recherche paramétrique sur la pondération des classes, ainsi que par le choix des métriques d'évaluation.

Nous avons observé, à l'aune des résultats de ces deux modèles, un progrès par rapport à une *baseline* consistant à ne prendre qu'une seule des caractéristiques (la meilleure) sur le nom. Autrement dit, on peut supposer que nos modèles sont bien capables d'apprendre des informations fournies. On peut également déduire que les différentes caractéristiques apportent des informations diverses, complémentaires, dans la définition de la similarité de deux marques.

Concernant les modèles testés, on ne peut que noter que la forêt aléatoire obtient de meilleurs résultats que la régression logistique, que ce soit en terme de F1 score, ou analysant la courbe ROC, ce qui n'est pas surprenant étant données les capacités de ces deux classifieurs.

5.3 Conclusion sur les résultats

Ces premiers résultats montrent qu'il y a bien un intérêt à utiliser des méthodes d'apprentissage pour déterminer la similarité entre marques. Cependant, on peut également conclure qu'il est essentiel de revoir la façon d'évaluer ces méthodes, avec une plus grande interaction avec les experts. Le fait d'obtenir de bons résultats selon les métriques classiques ne garantit pas dans les faits une applicabilité du système développé, ni sa pérennité dans le temps.

Il semble primordial de prendre en compte ces problématiques avant d'envisager le développement de nouveaux systèmes d'apprentissage dans le traitement de cette tâche. Les nouvelles méthodes proposées se heurteraient à la même difficulté d'évaluation et d'appréciation des capacités effectives.

Cette nécessité de repenser les métriques d'évaluation, et d'améliorer la façon de modéliser les différents aspects des marques, met en évidence des problématiques complexes à l'interface entre statistiques et considérations industrielles.

6 Conclusion

Dans ce travail, nous avons présenté une nouvelle approche pour déterminer la similarité entre marques verbales, basée sur des méthodes d'apprentissage automatique. Les résultats obtenus sont encourageants, mais ouvrent la voie à de multiples questions et améliorations, tant sur les plans scientifiques qu'industriels.

Features Un travail sur les *features* a été amorcé mais peut encore être approfondi, sur tous les critères. Malgré le fait que l'approche naïve pour intégrer les classes de Nice n'a pas porté ses fruits, il n'est pas exclu qu'une méthode plus approfondie contribue à améliorer les performances du modèle. Un travail sur l'aspect distinctif de la marque peut également être considéré : celui-ci est souvent au début, améliorant les résultats des mesures accordant un poids supérieur au début du signe (soundex, weighted Levenstein). On peut toutefois affiner sa prise en compte, avec une mesure statistique sur le poids calculé par *tf idf* par exemple.

Système d'apprentissage Un meilleur score d'optimisation du système que la F1 pour éviter les faux négatifs, problème crucial dans de nombreux champs d'application de l'IA. Le système d'active learning est pour l'instant à l'état de prototype et nécessite davantage de travail pour être déployé et soumis aux experts. Leurs retours seront déterminants pour définir dans quelle direction les efforts doivent se concentrer.

Evaluation L'évaluation doit s'affiner pour mieux saisir les retours qualitatifs des experts. Réussir à définir à partir de combien d'exemples le système est suffisamment fiable et surtout, capable de généraliser à de nouvelles marques présentant potentiellement d'autres caractéristiques. Par

ailleurs, le déploiement du système sur le long terme mérite une évaluation continue pour éviter des écueils comme le *data drift*.

Références

- [1] F. M. Anuar, R. Setchi, and Y. Lai. Semantic retrieval of trademarks based on conceptual similarity. *IEEE Transactions on Systems, Man, and Cybernetics : Systems*, 46(2) :220–233, 2016.
- [2] Leonidas Aristodemou and Frank Tietze. The state-of-the-art on intellectual property analytics (ipa) : A literature review on artificial intelligence, machine learning and deep learning methods for analysing intellectual property (ip) data. *World Patent Information*, 55 :37–51, 2018. Advanced Analytics of Intellectual Property Information for TechMining.
- [3] Kyung Pyo Ko, Kwang Hee Lee, Mi So Jang, and Gun Hong Park. 2-gram-based phonetic feature generation for convolutional neural network in assessment of trademark similarity. *CoRR*, abs/1802.03581, 2018.
- [4] Fatahiah Mohd Anuar, Rossitza Setchi, and Yu-Kun Lai. Trademark retrieval based on phonetic similarity. In *2014 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 1642–1647, 2014.
- [5] Taoxin Peng, Lin Li, and Jessie Kennedy. A comparison of techniques for name matching. *GSTF Journal on Computing (JoC)*, 2(1), 2014.
- [6] Rossitza Setchi and Fatahiah Mohd Anuar. Multi-faceted assessment of trademark similarity. *Expert Syst. Appl.*, 65 :16–27, 2016.
- [7] Taffee T Tanimoto. Ibm internal report. *Nov*, 17 :1957, 1957.
- [8] C. Trappey, A. Trappey, and Sam C. Lin. Intelligent trademark similarity analysis of image, spelling, and phonetic features using machine learning methodologies. *Adv. Eng. Informatics*, 45 :101120, 2020.
- [9] Wipo. *World intellectual property indicators 2021*. World Intellectual Property Organization, November 2021.