TU WIEN Informatics

TrademarkML

Die Vorhersage der Verwechslungsgefahr gemäß Artikel 8 Absatz 1 UMV

DIPLOMARBEIT

zur Erlangung des akademischen Grades

Diplom-Ingenieur

im Rahmen des Studiums

Data Science

eingereicht von

Maximilian Haller, BSc
Matrikelnummer 11810429

an der Fakultät für Informatik

der Technischen Universität Wien

Betreuung: Ao. Univ.Prof. Dr.iur. Markus Haslinger

Wien, 15. November 2023

Maximilian Haller	Markus Haslinger

TU WIEN Informatics

TrademarkML

Predicting the Likelihood of Confusion under Article 8(1) EUTMR

DIPLOMA THESIS

submitted in partial fulfillment of the requirements for the degree of

Diplom-Ingenieur

in

Data Science

by

Maximilian Haller, BSc
Registration Number 11810429

to the Faculty of Informatics

at the TU Wien

Advisor: Ao. Univ.Prof. Dr.iur. Markus Haslinger

Vienna, 15th November, 2023

Maximilian Haller	Markus Haslinger

Erklärung zur Verfassung der Arbeit

Maximilian Haller, BSc

Hiermit erkläre ich, dass ich diese Arbeit selbständig verfasst habe, dass ich die verwendeten Quellen und Hilfsmittel vollständig angegeben habe und dass ich die Stellen der Arbeit – einschließlich Tabellen, Karten und Abbildungen –, die anderen Werken oder dem Internet im Wortlaut oder dem Sinn nach entnommen sind, auf jeden Fall unter Angabe der Quelle als Entlehnung kenntlich gemacht habe.

Wien, 15. November 2023

Maximilian Haller

Danksagung

Ich möchte mich in erster Linie bei meinem Betreuer, Herrn Professor Haslinger, für die hervorragende Zusammenarbeit bedanken. Ihre Ratschläge waren und sind nicht nur fachlich, sondern auch persönlich, für mich von großem Wert. Ohne Ihr Vertrauen in mich und meine Arbeit, hätte ich mich nicht auf diese Querschnittsmaterie spezialisieren können. Selbst in den stressigsten Zeiten hatten immer ein offenes Ohr für meine Anliegen. Ich weiß das zu schätzen und bin Ihnen zutiefst dankbar.

Meine Dankbarkeit gilt auch meiner geliebten Familie, die mich immerzu unterstützt und aufgebaut hat. Auch wenn ich es nicht immer schaffe meinen Gefühlen entsprechenden Ausdruck zu verleihen, seid ihr alles für mich. Leider kann ich nicht an jedem Ort der Welt gleichzeitig sein.

Kurzfassung

In den letzten Jahren ist die Zahl der Markenanmeldungen im europäischen Raum kontinuierlich gestiegen. Um die Markenüberwachung für Markeninhaber oder die Markenrecherche für Antragsteller zu erleichtern, braucht es ein zuverlässiges Instrument zur Vorhersage von Verwechslungsgefahr gemäß Artikel 8 Absatz 1 UMV. Dies setzt voraus, dass nicht nur die Marken, sondern auch die entsprechenden Waren und Dienstleistungen, im Rahmen eines beweglichen Systems ähnlich sind. Derzeitige Systeme, die dem Vergleich von Marken dienen, sind für diese Vorhersage unzureichend, da diese oft nur den Markennamen berücksichtigen.

Aus diesem Grund wird in dieser Arbeit das Konzept eines Trademark-Management-Systems, TrademarkML, vorgestellt. Dieses soll die Markenüberwachung und Markenrecherche durch automatisierte Vorhersagen erleichtern.

Um diese Vorhersagen zu treffen, werden lege artis Methoden zur Extraktion von Merkmalen und zur Berechnung der Ähnlichkeit von Marken eingesetzt. Ein erschöpfender Vergleich zwischen der Brauchbarkeit dieser Merkmale wird im Rahmen dieser Arbeit durchgeführt, indem Random Forests und Support Vector Machines auf allen möglichen Merkmalskombinationen optimiert werden.

Durch diese Methode erreicht TrademarkML einen F1-score von 88% für Wortmarken und 81% für Bildmarken, einen Recall von 98% für Wortmarken und 84% für Bildmarken und eine Precision von 80% für Wortmarken und 77% für Bildmarken. Die Vorhersagen für Wortmarken erreichen im Durchschnitt einen um 6.8% höheren Wert für jede Metrik.

Keines der optimierten Modelle nutzt das Merkmal, das angibt, wie ähnlich sich die Waren und Dienstleistungen von zwei Marken sind. Daher ist TrademarkML nicht in der Lage, den Ausgang von teilweise zurückgewiesenen Widersprüchen korrekt vorherzusagen. Ein größerer Datensatz und weitere Methoden zur Extraktion von Merkmalen sind daher nötig, um dieses Problem zu bewältigen.

Abstract

Over the past years, the number of trademark application has risen continuously. To make it easier for trademark owners to protect their trademarks' territory and for trademark applicants to know if their trademark is likely to be refused, an automated and reliable tool is needed to compare two trademarks and their respective goods and services. Most of the existing services offering a similarity search just consider the spelling of the trademarks names, which is insufficient to assess likelihood of confusion.

For this reason, this thesis introduces the concept of a trademark management system, TrademarkML. TrademarkML faciliates tasks like trademark monitoring and searching for conflicting trademarks by automatically classifying trademark pairs.

TrademarkML employs state-of-the-art methods for extracting meaningful features for the comparison of trademarks in various aspects. Exhaustive feature selection is then used to tune random forests and support vector machines to all feature combinations.

TrademarkML achieves an F1-score of 88% for word marks and 81% for figurative marks, a recall of 98% for word marks and 84% for figurative marks, and a precision of 80% for word marks and 77% for figurative marks. Overall, TrademarkML performs better on word mark data by 6.8% on average for each metric.

None of the top-performing models uses a feature for measuring the similarity of goods and services. This makes it impossible to correctly predict partially upheld oppositions. A larger dataset and more meaningful features are required to overcome this issue.

Contents

CHAPTER 1

Introduction

1.1 Motivation and Problem Statement

Trademarks are a type of intellectual property that refers to recognizable symbols that identify and set apart goods and services from different origins. They act as indicators of origin for goods and services. To serve their purpose, trademarks must be distinguishable. Therefore, Article 8(1) EUTMR enables the proprietor of an earlier right to oppose the registration of later trademark applications in cases of likelihood of confusion and double identity.

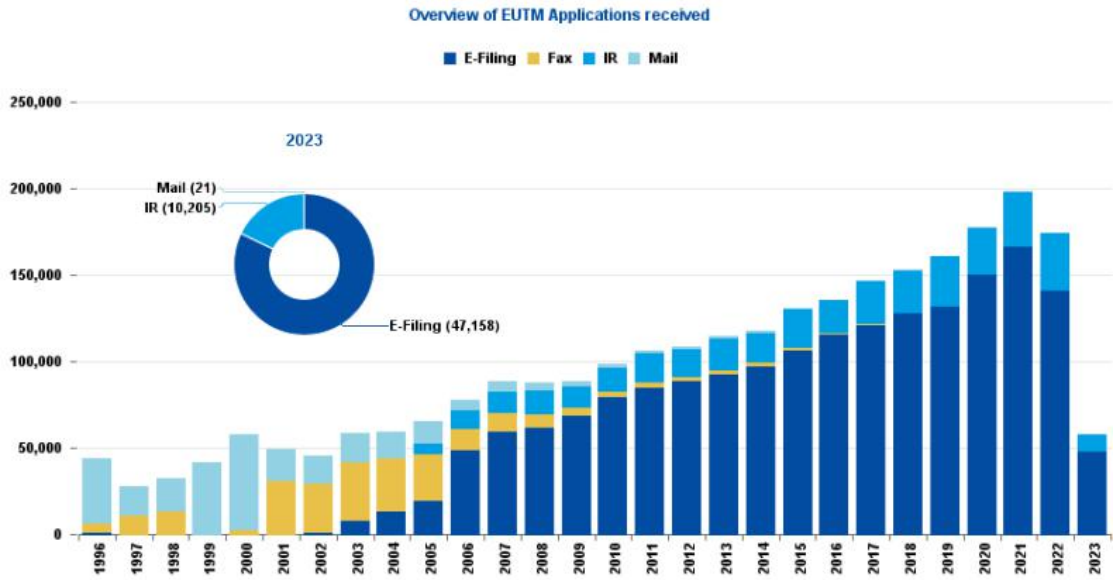


Figure 1.1: Number of EUTM Applications per Year [42, p 5]

The European Union Intellectual Property Office (EUIPO), formerly known as Office for Harmonization in the Internal Market (OHIM), does not assess likelihood of confusion eo ipso. Instead, the proprietor of the earlier right has to file an opposition against the applicant of the new trademark within a period of three months following the publication of the application in the so-called European Union Trade Marks Bulletin [44, p 734]. This leads to legal uncertainty for the applicant, as they cannot be sure whether their registration will succeed or not. Furthermore, trademark owners are confronted with constantly rising numbers of trademark applications, as can be seen in figure 1.1, which makes it more and more uneconomical to manually check for potential trademark infringement.

To make it easier for trademark owners to protect their trademark's territory and for trademark applicants to know if their trademark is likely to be refused, an automated and reliable tool is needed to compare two trademarks and their respective goods and services. Most of the existing services offering a similarity search just consider the spelling of the trademarks' names, which is insufficient to assess likelihood of confusion.

The goal of this thesis is to provide a machine-learning pipeline, TrademarkML, that allows to predict the likelihood of confusion for two word trademarks or two figurative trademarks. This makes it possible to automatically process a large set of trademarks like the ones found in the European Union Trade Marks Bulletin or the European trademark database [40].

1.2 Research Questions

This diploma thesis aims to answer the following questions:

1. **What features relevant to the evaluation of likelihood of confusion under Article 8(1) EUTMR can be computed from an opposition case?**
   This question aims to identify ways to compute meaningful features for the assessment of the likelihood of confusion. Features to be computed are mainly similarity scores that are derived from so-called similarity factors. In order to identify these features, existing literature is studied.

2. **Which prediction performance is achieved by TrademarkML?**
   Within the context of this question a machine-learning pipeline with several configurations is developed and evaluated. Answering this question also includes selecting features identified in RQ1. The evaluation is performed against data taken from previous opposition cases.

1.3 Methodology

In order to answer the research questions defined above, the following steps are carried out:

1. **Legal Analysis**: In this step, legal provisions, opposition decisions by the EUIPO, examination guidelines by the EUIPO, and court decisions by the Court of Justice of the European Union (CJEU) are analyzed. The main goal is not to look at each single instance of case law but rather to quantify rulings and to identify factors that are relevant for the outcome of an opposition case based on Article 8(1) EUTMR.

2. **Systematic Literature Review**: A systematic literature review is performed following the guidelines developed by Kitchenham and Charters [81]. The literature review serves as theoretical foundation of the implementation. The goal is to find methods to extract features that simulate the factors identified in the legal analysis. The protocol for this review can be found in appendix 9.

3. **Dataset Creation**: Since previous EUIPO opposition decisions are unstructured data that cannot be used out of the box, it is necessary to manually create a subset of data to automatically process it. For this reason, instances from the trademark case law database [39] as of 25 August 2023 are used. The dataset consists of 500 opposition cases. It is balanced and consists of 250 cases dealing with word marks and 250 cases dealing with figurative marks. For each of these types of marks, the class label must also be balanced. Cases must have the following characteristics to be included in the dataset:

    - Both trademarks must have a name.
    - Both trademarks must be of the same type.
    - The decision must be based on the assessment of likelihood of confusion or double identity.
    - The decision must be in English.
    - The decision must only consider two trademarks.
    - The application must be admissible.
    - For figurative marks, images for both marks must be available.

4. **Implementation**: In this step, the prediction module of the trademark management system, TrademarkML, is created. This is done using the theoretical knowledge and state-of-the-art methods found in the literature review. For problems that have not yet been addressed by the literature, novel approaches are developed and current state-of-the-art methods are adapted to the use case of this thesis. The goal is to create a pipeline that allows for extensive evaluation by accepting different configurations.

5. **Evaluation**: As a last step, the pipeline is tested with different configurations. The models performance is evaluated and used for a qualitative and quantitative analysis. Since ground truth is available for each case, the model is solely tested against the ground truth. Lastly, potential limitations and biases are investigated and discussed. The evaluation includes inter alia recall, precision, and accuracy.

CHAPTER 2

Legal Analysis

In this chapter, relevant factors for examining the likelihood of confusion are identified. This is done by investigating legal foundations, guidelines for examination, and previous opposition decisions.

2.1 Legal Foundations

Trademark law is part of intellectual property law and deals with the rights and obligations of trademark owners, the requirements and limitations for trademarks, and the procedures in the context of trademark law. The scope of this thesis is European trademark law, which is part of European intellectual property law. In contrast to each member state's national trademark law, European trademark law offers extensive protection for a trademark in all current and future European Union (EU) member states [41]. Relevant legal texts concerning European trademark law and its procedures are the EUTMR, the EUTMDR, and the EUTMIR [45]. Furthermore, the Paris Convention contains important provisions for industrial property that also have to be considered in Euopean trademark law [44, p 781].

The EUTMR came into effect on the first of October 2017. The main changes were the introduction of EU certification marks and the elimination of the requirement of graphical representation for trademarks [47, p 6]. The EUTMDR contains rules for procedures related to European Union trademarks and the EUTMIR covers details to the procedure, like the contents of EUTM applications [47, p 6].

Article 8 EUTMR lists the so-called relative grounds for refusal. Relative grounds for refusal are only examined upon opposition and are not, in contrast to absolute grounds for refusal, examined ex officio. Both likelihood of confusion and double identity are relative grounds for refusal and defined in Article 8(1) EUTMR.

Article 2 EUTMR provides for an institution which is responsible for the implementation of European trademark law. The EUIPO, often referred to as "the office", existed already before the EUTMR came into effect, but was formerly known as OHIM. According to recital 2 of Regulation (EU) 2015/2424 its name was changed due to the Lisbon Treaty coming into effect. The EUIPO carries out legal and administrative procedures required for implementing European trademark law and law concerning community designs autonomously. Therefore, it is also responsible for the assessment of likelihood of confusion. The guidelines for examination used by the EUIPO [44] are the main source of information for finding relevant factors that determine likelihood of confusion.

2.2 Concept of Trademarks

The legal definition of a trademark depends on the respective legal system. As this thesis is written in the context of European intellectual property law, only the definition for European Union Trademarks (EUTMs) is relevant.

The legal definition of EUTMs can be found in Article 4 EUTMR.

> **Article 4 EUTMR**
>
> An EU trade mark may consist of any signs, in particular words, including personal names, or designs, letters, numerals, colours, the shape of goods or of the packaging of goods, or sounds, provided that such signs are capable of:
>
> (a) distinguishing the goods or services of one undertaking from those of other undertakings; and
>
> (b) being represented on the Register of European Union trade marks (the Register), in a manner which enables the competent authorities and the public to determine the clear and precise subject matter of the protection afforded to its proprietor.

2.2.1 Types

The use of the words "in particular" implies that the list of Article 4 EUTMR is not exhaustive, meaning that the scope of trademarks is not limited to the signs explicitly mentioned. This broad definition allows for various types of trademark, like word marks, figurative marks, figurative marks containing word elements, shape marks, shape marks containing word elements, position marks, pattern marks, single color marks, color combination marks, sound marks, motion marks, multimedia marks and hologram marks.

One question that arises from this definition is if olfactory marks can be trademarks under European trademark law. CTM 000428870 is the only olfactory mark that has ever existed in the EUTM database [40]. This mark, however, expired in 2006. In

Type	Absolute #Trademarks	Relative #Trademarks	Absolute #Oppositions	Relative #Oppositions
Word	1.517.714	57.094%	58.260	57.746%
Figurative	1.124.922	42.318%	42.009	41.638%
3D Shape	11.885	0.447%	144	0.142%
Other	1205	0.045%	447	0.443%
Color	1126	0.042%	26	0.025%
Sound	477	0.018%	0	0%
Position	422	0.015%	3	0.0003%
Motion	216	0.008%	0	0%
Pattern	186	0.007%	1	0.0099%
Multimedia	90	0.003%	0	0%
Hologram	13	0.000%	0	0%
Olfactory	0	0.000%	0	0%

Table 2.1: Number of Trademarks and Opposition Cases per Type according to [39] and [40] as of 25 August 2023

its judgements[1], the CJEU clarified that olfactory marks cannot be trademarks under European trademark law as olfactory marks cannot be represented in a way to comply with Article 4(b) EUTMR.

The only types relevant to this thesis are word marks, figurative marks and figurative marks containing word elements, as they make up the vast majority of trademarks as can be seen in table 2.1. Not only are more than 99% of the trademarks of type word or figurative, but also 98% of all oppositions are based on these two types of trademarks.

Word Marks

Article 3(3)(a) EUTMIR defines word marks as marks that consist exclusively of words, letters, numerals, or other standard typographic characters or a combination thereof. In other words, word marks are marks that can be represented solely as text. Word marks do not have any color or additional graphical features.

The examples in table 2.2 demonstrate that word marks are restricted to the textual representation. However, within that textual form, there is no restriction to the marks' syntax.

[1]Judgement of 12 December 2002, C-273/00, *Ralf Sieckmann v Deutsches Patent- und Markenamt*, EU:C:2002:748 and Judgement of 27 October 2005, T-305/04, *Eden SARL v European Union Intellectual Property Office*, EU:T:2005:380.

Trademark ID	Mark
EUTM 002288355	adidas
EUTM 002890218	SPACEX
IR W01214710	#tag
EUTM 018814133	Thanks milk. We'll take it from here.

Table 2.2: Examples of Word Marks

Figurative Marks

In contrast to word marks, Article 3(3)(b) EUTMIR defines figurative marks as marks that contain additional graphical features or nonstandard characters. European law does not differentiate between figurative marks and combined marks. A combined mark is therefore just a special case of figurative mark that contains verbal elements. In table 2.3, all marks except for EUTM 000002337 are combined marks.

Trademark ID	Mark
EUTM 000001157	
EUTM 000002337	
EUTM 000053819	
EUTM 000071142	

Table 2.3: Examples of Figurative Marks

2.2.2 Geographical Scope of Protection

The scope of protection depends on the system under which a trademark is registered. There are national, regional, EU-wide, and international systems [43]. A EUTM is a trademark registered in the regional trademark system of the EU. This system is grounded on the EUTMR. According to recital 7 EUTMR, the EUTM-system is built on top of the laws of the member states, meaning that each member state still has their own national trademark system. However, the unitary EUTM-system is the only trademark system that provides protection for the entire territory of the EU [56, p 220 f]. National level trademarks are useful if a protection at EU-level is not wanted or required [43].

2.2.3 Grounds for Refusal

According to recital 11 EUTMR, the main purpose of trademarks is to indicate the origin of a good or service. However, not every sign is recognized as a valid trademark by European trademark law. In order to be registrable, a mark must not fail on relative nor absolute grounds for refusal. While absolute grounds of refusal are to be examined upon application, relative grounds of refusal require a preceding admissible opposition filed by the proprietor of an earlier trademark or other form of trade sign [44, p 736]. The goal of this section is to provide an overview of grounds for refusal, since they provide a negative definition for registered trademarks.

Absolute Grounds for Refusal

Article 7(1)(a) EUTMR ensures that trademarks comply with the requirements of Article 4 EUTMR. In other words, the sign itself must be capable of distinguishing goods and services from different origins and it must be representable on the Register [44, p 363]. The CJEU made clear that abstract concepts and ideas are not sufficiently specific to be a sign under Article 4 EUTMR.[2] Article 4(a) EUTMR also requires the sign to have the abstract capacity to distinguish goods and services of different origins [44, p 364] [56, p 221]. Furthermore, Article 4(b) EUTMR requires signs to be capable of being represented on the Register. The incapability of being represented on the Register affects non-traditional marks, like olfactory marks [56, p 225 f].[3] Article 7(1)(b) EUTMR states that trademarks must be distinctive. However, according to case-law a minimum degree of distinctiveness is sufficient to prevent application of Article 7(1)(b) EUTMR.[4] Distinctiveness is the attribute that indicates that the trademark is suitable for identifying goods and services from a certain origin [44, p 381] [132, p 15]. According to the CJEU[5], lexical structures common in advertising language are considered non-distinctive.

Article 7(1)(c) EUTMR makes sure that trademarks are non-descriptive. A trademark is considered descriptive if its sign provides information about characteristics of the goods or services for which the registration is sought [44, 468]. In case the mark is descriptive, it is also considered as non-distinctive [44, p 468]. Article 7(1)(d) EUTMR ensures that trademarks do not consist exclusively of signs that have become customary in the current language. In contrast to Article 7(1)(c) EUTMR, this norm is relevant for cases where the original meaning of a word has no direct relationship to the good or service [44, p

[2] Judgement of 21 April 2010, T-7/09, *Schunk GmbH & Co. KG Spann- und Greiftechnik v OHIM*, EU:T:2010:153, § 25.

[3] Judgement of 12 December 2002, C-273/00, Ralf Sieckmann v Deutsches Patent- und Markenamt, EU:C:2002:748 and Judgement of 27 October 2005, T-305/04, Eden SARL v European Union Intellectual Property Office, EU:T:2005:380.

[4] Judgement of 3 April 2019, T555/18, *Medrobotics Corp. v EUIPO*, EU:T:2019:213, § 19 and Judgement of 28 June 2017, T-479/16, *Colgate-Palmolive Co. v European Union Intellectual Property Office*, EU:T:2017:441, § 19 and Judgement of 25 September 2015, T-591/14, *BSH v OHIM*, EU:T:2015:700, § 40.

[5] Judgement of 25 April 2013, T-145/12, *Bayerische Motoren Werke AG v OHIM*, EU:T:2013:220, ğ 29.

524]. Article 7(1)(e) EUTMR says that trademarks must not consist exclusively of a shape or characteristic that results from the nature of the good itself or from its value. Article 7(1)(d) EUTMR is not restricted to 3D shapes and is also applicable to figurative marks [44, 529].

Article 7(1)(f) EUTMR guarantees that trademarks comply with public policy and acceptable principles of morality. This norm is related to Article $6^{quinquies}$(B)(3) of the Paris Convention [44, p 544]. Whereas the public policy is a concept based on objective criteria like principles and fundamental values of the EU [44, p 546], acceptable principles of morality refer to subjective values that are likely to change over time [37, p 536 f].[6] Article 7(1)(g) EUTMR prevents the registration of marks with a deceptive character. The deceptiveness of a mark depends on the goods and services the mark shall be registered for [44, p 561]. However, deceptiveness can only exist if the mark is sufficiently specific, which requires the mark to clearly indicate characteristics of the respective goods and service.[7] Article 7(1)(h) EUTMR is related to Article 6^{ter} Paris Convention and protects symbols of states that are party to the Paris Convention by refusing the registration of marks that are identical to them [44, p 588]. Furthermore, Article 7(1)(i)EUTMR protects additional emblems that are not covered by Article 6^{ter} Paris Convention.

Article 7(1)(j) EUTMR prevents the registration of marks that conflict with Geographical Indications (GIs) that enjoy protection in the EU. In the EU, wines are protected under Regulation (EU) 1308/2013, spirit drinks are protected under Regulation (EU) 2019/787, and agricultural products and foodstuffs are protected under Regulation (EU) 1151/2012. GIs can also be protected through international agreements [44, p 609]. Article 7(1)(k) EUTMR provides for the refusal of applications in case their marks conflict with Traditional Terms of Wine (TTWs) which are also protected under Regulation (EU) 1308/2013.

Article 7(1)(m) EUTMR is applicable in cases where essential elements of a trademark conflict with a registered Plant Variety Denomination (PVD). However, this is assessed only when the trademark registration is sought for live plants, agricultural seeds, fresh fruits, or fresh vegetables [44, p 685].

Article 7(3) EUTMR provides for the registration of a trademark regardless of the applicability of Article 7(1)(b)-(d) EUTMR in case a mark has acquired distinctiveness through use. This possibility is only examined upon request by the applicant [44, p 694].

Relative Grounds for Refusal

Relative grounds for refusal consider the relation between two trademarks [36, p 294] and are only examined if a proprietor of an earlier trademark or other form of trade

[6] Judgement of 27 February 2020, C-240/18 P, *Constantin Film Produktion GmbH* v *EUIPO*, EU:C:2020:118, § 39.

[7] Judgement of 24 September 2008, T-248/05, *HUP Uslugi Polska sp. z o.o.* v *OHIM*, ECLI:EU:T:2008:396, § 65 f and Judgement of 29 November 2018, T-681/17, *Khadi and Village Industries Commission* v *EUIPO*, EU:T:2018:858, § 53.

sign files an admissible opposition against a trademark [44, p 736]. Relative grounds for refusal relevant to this thesis are likelihood of confusion and double identity, which are constituted in Article 8(1) EUTMR.

2.3 Concept of Likelihood of Confusion and Double Identity

Likelihood of confusion and double identity are defined in Article 8(1) EUTMR.

> **Article 8(1) EUTMR**
>
> Upon opposition by the proprietor of an earlier trade mark, the trade mark applied for shall not be registered:
>
> (a) if it is identical with the earlier trade mark and the goods or services for which registration is applied for are identical with the goods or services for which the earlier trade mark is protected;
>
> (b) if, because of its identity with, or similarity to, the earlier trade mark and the identity or similarity of the goods or services covered by the trade marks there exists a likelihood of confusion on the part of the public in the territory in which the earlier trade mark is protected; the likelihood of confusion includes the likelihood of association with the earlier trade mark.

While Article 8(1)(a) EUTMR provides for the opposition in case of double identity, Article 8(1)(b) EUTMR provides for the opposition in case of likelihood of confusion. The difference is that if both, the trademarks and the respective goods or services, are identical, there is no need to carry out an evaluation of likelihood of confusion [44, p 868]. In the context of this thesis, it is not important to differentiate between double identity and likelihood of confusion, as double identity can be seen as a special case of likelihood of confusion where trademark similarity and the similarity of the respective goods or services are close to 100%. On the other hand, rejecting an opposition based on Article 8(1)(a) EUTMR does not imply that the opposition would also fail on the ground of Article 8(1)(b) EUTMR.

According to [52, p 317 f], the CJEU identified the relevant criteria for assessing likelihood of confusion between marks and goods and services in the cases *Sabel v Puma*[8] and *Canon*[9]. Regarding the similarity of two marks visual similarity, aural similarity, conceptual similarity, and the inherent or acquired distinctiveness of the earlier mark are relevant factors. Furthermore, the degree of attention paid by the relevant public plays a role in the assessment of likelihood of confusion [44, p 946]. The similarity of goods and

[8] Judgement of 11 November 1997, C-251/95, *SABEL BV* v *Puma AG*, EU:C:1997:528.

[9] Judgement of 29 September 1998, C-39/97, *Canon Kabushiki Kaisha* v *Metro-Goldwyn-Mayer Inc.*, EU:C:1998:442.

services is assessed by taking into account their nature, intended purpose, method of use, and the connection between them. However, this list of factors is non-exhaustive [52, p 317]. The similarity between marks and the similarity between goods and services are interdependent.[10]

To carry out legal and administrative procedures in a standardized way, the EUIPO has defined examination guidelines [44]. Part C section 2 of these guidelines [44, p 863 ff] is the relevant source of information on how comparisons between marks or goods and services are made.

2.3.1 Similarity of two Marks

When assessing the similarity of two marks under Article 8(1) EUTMR, signs must be compared visually, aurally, and conceptually [44, p 958]. For each aspect, the degree of similarity must be examined by comparing the signs in their entirety [44, p 959]. However, the comparison can be restricted in case of negligible elements [44, p 960]. In the following sections, an overview is given on what these similarities are. Extensive details on case-law are omitted.

Visual Similarity

The visual comparison of two word marks is different to the visual comparison of two figurative marks, as word marks do not contain additional figurative elements. While usually it does not matter whether word marks are written in lower case or upper case, irregular capitalization can have an impact on the visual similarity of two word marks.[11] The examination guidelines do not comment on what aspects of the words contribute to the assessment in which way. However, the CJEU made clear that the presence of a sequence of characters in both word marks is essential to visual similarity.[12]

Table 2.4 contains samples for every degree of visual similarity. The following paragraphs outline the arguments considered by the EUIPO when asserting the degree of visual similarity in the respective opposition decisions.

While the visual similarity of “PREDATOR” and “Predator” can be easily derived by comparing the character sequences in a case insensitive way, the reasons for the other examined degrees of visual similarity in the rest of the samples are not obvious at first sight.

This is because the EUIPO Opposition Division takes into account the degree of distinctiveness of character sequences when comparing word marks. For the marks “DRIP

[10] Judgement of 29 September 1998, C-39/97, *Canon Kabushiki Kaisha* v *Metro-Goldwyn-Mayer Inc.*, EU:C:1998:442, § 17.

[11] Opposition Decision of 31 March 2016, *AIDA Cruises* v *Damia GmbH*, R 3290/2014-4, § 38 in contrast to Judgment of 31 January 2013, *Present-Service Ullrich* v *OHIM*, T66/11, not published, EU:T:2013:48, § 57 f and Judgement of 27 January 2019, *REWE-Zentral AG* v *OHIM*, T-331/08, EU:T:2010:23, § 16 f and Judgement of 11 June 2014, *Sofia Golam* v *OHIM*, T281/13, EU:T:2014:440, § 41.

[12] Order of 4 March 2010, *Kaul GmbH* v *OHIM*, C-193/09 P, EU:C:2010:121, § 83.

Case ID	Sign 1	Sign 2	Visual Similarity
003158578	PREDATOR	Predator	identical
003163445	DRIP DROPS CBD	DRIP DROP	high
003166793	PATERICO	MATERICO	average to high
003177425	Volkspflege	VOLKSWAGEN	average
003112688	WE 11 DONE	WE	low to average
003150360	ki-Tec	KITE	low

Table 2.4: Examples of Visual Similarity between Word Marks

DROPS CBD" and "DRIP DROP" the EUIPO Opposition Division argues that the substring "DRIP DROP" is inherently distinctive while "CBD" has a weak distinctive character as this character sequence is associated with a substance that can be found in all opposed goods. As the strings are not identical but both marks share the same distinctive character sequence, their visual appearance is considered highly similar.

"PATERICO" and "MATERICO" are also almost identical strings, but their degree of visual similarity is considered average to high. However, these two strings are both distinctive and one mark does not contain the other mark. Thus, they have a slightly lower degree of visual similarity. "Volkspflege" and "VOLKSWAGEN" differ in three letters that are not visually separated from the distinctive elements of the two marks. While the dominant character sequence "VOLKS" is contained in both strings, the rest of both strings shares only the letters "GE". The degree of visual similarity is considered to be average. The word mark "WE" is contained in the opponent's mark "WE 11 DONE", similar to case number 003163445. However, none of the elements in these marks are highly distinctive. The character sequence "WE" is considered to have a weak distinctive character. As both marks start with this sequence, they are considered to be visually similar to a low to average degree.

Lastly, "ki-Tec" and "KITE" are visually similar to a low degree. The EUIPO Opposition Division argues that the irregular capitalization, the presence of the hyphen, and the addition of the letter "c" at the end of "ki-Tec" lead to perceptible differences in the overall impressions. Opposition decision 003150360 also comments on arguments that were brought forward by the parties. The EUIPO Opposition Division approves that, generally, the beginning of a sign has a rather strong impact on the consumer's perception, and that small differences have a stronger impact on the perception of short signs.

In contrast to the comparison of word marks, the comparison of figurative marks considers the marks' stylization [44, p 988 f].

The comparison of figurative marks is more complex. Measuring the similarity or the distance between two figurative marks must consider not only word elements contained in the figures but also the signs' structures, stylizations, colors, and contours. Furthermore,

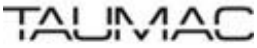
Case ID	Sign 1	Sign 2	Visual Similarity
002173717			identical
003176608			high
003174765			average to high
003173691			average
003138449			average to low
003167224			low
003152689			dissimilar

Table 2.5: Examples visual similarity between figurative marks

signs can be considered visually identical even if they are technically not identical (see case number 002173717 in table 2.5). Even though figurative elements are classified under the Vienna Classification, the signs' classification does not influence the examination of visual similarity, as the examination is based solely on the signs.

Aural Similarity

The aural similarity is derived by comparing two marks' overall phonetic impressions. The phonetic impression comes from the syllables, the sequence of syllables, and their respective prominence [44, p 1004]. There is a certain interdependence between aural and visual similarity, as words that sound similar will usually also have a similar spelling. This effect can be seen in figure 4.4. Since purely figurative marks cannot be pronounced, their aural similarity cannot be assessed [44, p 1006].

Table 2.6 contains samples for every degree of aural similarity. The following paragraph outlines the arguments considered by the EUIPO when assessing the degree of aural similarity in the respective opposition decisions.

Although there is a certain interdependence between aural and visual similarity, they compare different aspects of the signs. This can be seen in cases 003170293 and 003160216. While the spelling of "LIFE'S" and the spelling of "LIVE" are quite different, the only difference in their pronunciation is the tailing "s". Also, "DREAMIN*101'" has a

Case ID	Sign 1	Sign 2	Aural Similarity
003170820	IDA	IDA	identical
003170293	LIFE'S	LIVE	high
003160216	DREAMIN*101'	DREAMS	average to high
003176903	EMBACO	EMBACOLLAGE	average
003153372	LÜTZE CABLEFIX	CABLEFIX	low to average
003174830	SWP PROCOR	PROCORALAN	low
003146419	HOME deco	MGI H HOME DECÓ	dissimilar

Table 2.6: Examples of aural similarity between marks

completely different visual appearance than "DREAMS". However, the only part of the first sign, that is actually pronounced, is just "DREAMIN". Thus, "DREAMIN*101'" and "DREAMS" are aurally similar to an average to high degree. The aural similarity of all other samples can be derived from rules already known from the section about visual similarity 2.3.1.

Conceptual Similarity

Conceptual similarity is a concept that heavily differs from visual and aural similarity, as it does not compare the signs themselves but their concepts instead. This, however, requires the signs to evoke a concept. Since not all signs are related to a concept, the conceptual comparison plays a varying role in the assessment of likelihood of confusion [44, p 1049]. The EUIPO equates concepts with semantic meaning. While there is no clear definition for these terms, the EUIPO makes clear that the conceptual relationship is not relevant for broad categories. The signs "PEAR" and "APPLE BITE" might be conceptually related fruits, but their common features related to the shared concept have a very limited impact on the overall impression.[13] From a practical point of view, the conceptual similarity can often not be assessed as can be seen in Figure 4.3 and therefore plays a subordinate role in the assessment of likelihood of confusion.

Table 2.7 contains examples for every degree of conceptual similarity. The following paragraph outlines the arguments considered when asserting the degree of conceptual similarity in the respective opposition decisions.

"MOVEUP" and "moveUP" are considered conceptually identical as both marks will be identified with the phrasal verb "move up". "SHAMAN" and "SHAMAN'S" are both related to the concept of a shaman. However, due to the tailing "s" the meaning of one mark is slightly altered. Therefore, the signs are conceptually similar to a high degree but not identical. "Nonna Filomena" and "SANTA FILOMENA" are considered to share

[13] Judgement of 31 January 2019, T-215/17, *Pear Technologies Ltd* v *EUIPO*, ECLI:EU:T:2019:45, §§ 77-79.

Case ID	Sign 1	Sign 2	Conceptual Similarity
003163877	MOVEUP	moveUP	identical
003179493	SHAMAN	SHAMAN'S	high
003173672	Nonna Filomena	SANTA FILOMENA	average to high
003180340	NOWGROW	NOWGO	average
003159802	LOVE VEGE	végé	low to average
003170087	Fresh Up	I-FRESH DÁVI	low
003178074	TP HOME	TP	dissimilar

Table 2.7: Examples of conceptual similarity between marks

a distinctive concept evoked by the female name "Filomena". "Nonna" and "SANTA", however, are considered meaningless. For this reason, these signs are conceptually similar to an average to high degree. "NOWGROW" and "NOWGO" are considered conceptually similar to an average degree as these expressions can be seen as a "call to action" in relation to financial services. The signs "LOVE VEGE" and "végé" both contain the substring "vege", which has been found to be related to the concept of vegetarianism.[14] For this reason, these two signs are conceptually similar to a low to average degree. "Fresh Up" and "I-FRESH DAVI" are both loosely related to concepts related to freshness. Therefore, they are conceptually similar to a low degree.

Distinctiveness

The distinctiveness of the earlier mark is taken into account in the context of the global assessment and is especially important when the signs are similar to only a low degree. A mark's distinctive character determines the strength and breadth of its protection [44, p 1103]. Distinctiveness is defined as the capacity to identify goods or services as coming from a particular undertaking [44, p 1103].

Earlier marks are presumed to be valid and, therefore, have a minimum degree of inherent distinctiveness. When a mark is related to characteristics of its goods and services, it is considered to have a low degree of distinctiveness except for when the allusion to characteristics is sufficiently imaginative [44, p 1104]. If there is no indication for a limited distinctiveness, the mark is considered to have a normal inherent distinctiveness [44, p 1104]. Earlier marks can also acquire a higher degree of distinctiveness. However, the presence of the prerequisites of this circumstance must be proven by submitting appropriate evidence [44, p 1104].

Table 2.8 lists three samples, each having a different degree of distinctiveness with regard to their respective goods and services. "MEDITERRANI" was found to only have low degree of distinctiveness, as its name alludes to the goods sugar, tea, and coffee.

[14] Judgement of 26 July 2023, T-434/22, *Topas* v *EUIPO*, ECLI:EU:T:2023:426, § 49.

Case ID	Earlier Mark	Goods and Services	Distinctiveness
003157085	TROPICO	Fruit juices	enhanced
003175010	APRANTA	Computer software	normal
003175766	MEDITERRANI	Sugar, tea, coffee	low

Table 2.8: Examples of degrees of distinctiveness

"TROPICO", on the other hand, provided evidence that it acquired an enhanced level of distinctiveness through extensive use of the mark.

However, from a practical point of view, in the majority of cases the distinctiveness is considered to be normal as can be seen in figure 4.3, which means that this attribute does not have a strong impact on the global assessment.

Degree of Attention

In the global assessment, the degree of attention of the relevant public is also taken into consideration. A higher degree of attention means that there is a smaller chance for the consumer to confuse goods and services of two origins and vice versa. The degree of attention depends on various factors, like the target consumer group and the nature of the goods and services [44, p 951].

The EUIPO lists three categories of goods and services where the degree of attention is generally considered to be high. These categories are expensive purchases, potentially hazardous purchases, cases of brand loyalty, and pharmaceuticals [44, p 952 f]. On the other hand, a lower degree of attention is assumed for categories of goods and services that are subject to habitual buying behavior [44, p 954].

2.3.2 Similarity of Goods and Services

The scope of protection of trademarks is defined by the goods and services for which the respective trademark is registered [57]. Thus, likelihood of confusion does not only require similar signs but also similar goods and services. This means that there might not be a likelihood of confusion even if the signs are identical.[15] The comparison of the goods and services is carried out ex officio and is limited to well-known facts, excluding knowledge of highly technical nature [44, p 895]. Many different factors are considered when assessing the similarity of goods and services. The nature, intended purpose, method of use, complementarity, and competition are the so-called "canon factors" according to the CJEU judgement *Canon*[16] [44, p 904]. However, there are also additional factors, like distribution channels, relevant public, and the usual origin of goods and services, that

[15] Opposition Decision of 27 February 2023, *Helsana Versicherungen AG* v *APW Consulting Agnieszka Pawowska-Wypych*, B 002756016.

[16] Judgement of 29 September 1998, C-39/97, *Canon Kabushiki Kaisha* v *Metro-Goldwyn-Mayer Inc.*, EU:C:1998:442.

play a role in the assessment of the similarity of goods and services [44, p 904]. The Nice Classification, which is a taxonomy for goods and services in the context of trademark law, serves purely administrative purposes and, therefore, cannot be used to directly infer the similarity between goods and services [44, p 884]. The Nice Classification may be used, however, to determine the nature and purpose of goods and services or to interpret the scope of protection [44, p 885 f].

It is important to note that goods and services may coincide in their wordings, but refer to different products. For example, "drills" in Class 7 of the Nice Classification refer to machine tools and are not identical to "drills" in Class 8, which refer to hand tools [44, p 897].

Goods and services are deemed to be identical if goods and services of one mark are contained in a broader category of goods and services of the other mark [44, p 898 f]. In many cases, even partially overlapping categories can be considered identical, if these goods and services cannot be clearly separated [44, p 900].

Since the list of factors used to compare goods and services is not exhaustive, a detailed analysis of case-law is omitted in this section.

2.3.3 Global Assessment

After assessing all relevant factors for likelihood of confusion, a global assessment has to be made. This final step incorporates the interdependence principle, meaning that a higher degree of similarity of goods and services can outweigh a lower degree of similarity of the marks and vice versa [44, p 1130]. Furthermore, different aspects of the signs' similarities can be considered more important than others. For example, the aural similarity could be more important for goods and services that are bought in noisy environments. The global assessment draws conclusions for each and every good or service in question. This means that for some goods and services the opposition can be rejected while for others the opposition is upheld.

CHAPTER 3

Background

The goal of this chapter is to provide the knowledge relevant for understanding the functionality and the evaluation of TrademarkML. Legal aspects and details specific to the implementation of TrademarkML are discussed in chapters 2 and 6.

3.1 Machine Learning

Machine learning is a part of artificial intelligence that is concerned with the development of models that learn relationships implicitly from the data, meaning that the connections between input and output must not be explicitly programmed. Many tasks can be solved with machine learning, like clustering data or predicting time series. However, in the context of this work, the focus lies on classification, which is a task addressed using supervised learning. Supervised learning is an approach where a computer program learns such a representation from experience [110, p 2].

This means that these models need to be trained with data, which consists of the target values and the respective features. The prediction performance does not only depend on the chosen algorithm but also on the given features. For this reason, feature engineering, which is an umbrella term for feature extraction, selection, and preprocessing, plays an important role when working with shallow architectures [119, p 1].

Formally, supervised machine learning models are trained and tested using a dataset $\boldsymbol{\mathcal{D}} = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_n, y_n)\}$ that contains n samples in which each tuple $(\boldsymbol{x}_i, y_i)$ consists of input data $\boldsymbol{x}_i \in \boldsymbol{X}$ and target value $y_i \in \boldsymbol{y}$. Since for every sample $\boldsymbol{x}_i$ just one value y_i exists, $\boldsymbol{y} \in \mathbb{R}^n$. Each $\boldsymbol{x}_i \in \boldsymbol{X}$ consists of p features $\boldsymbol{\mathcal{F}}^p = [\boldsymbol{X}_{:,1}, \boldsymbol{X}_{:,2}, \ldots \boldsymbol{X}_{:,p}]$ which means that $\boldsymbol{X} \in \mathbb{R}^{n \times p}$. The goal is to find function $f : \boldsymbol{X} \rightarrow \boldsymbol{y}$. Since the exact relationship between $\boldsymbol{X}$ and $\boldsymbol{Y}$ often cannot be found, it is approximated using $\hat{f} : \boldsymbol{X} \rightarrow \hat{\boldsymbol{y}}$ so that $\hat{\boldsymbol{y}} \approx \boldsymbol{y}$. However, a machine learning model is usually trained on a subset of $\boldsymbol{\mathcal{D}}$, the training data $\boldsymbol{\mathcal{D}}_{train}$, and then evaluated on test data $\boldsymbol{\mathcal{D}}_{test}$ unknown to

the trained model. This way, the evaluation considers the model's capacity to generalize to unseen data.

3.1.1 Classification

The goal of classification is to predict a class membership of a sample [102, p 3]. This means that in this task, $\boldsymbol{y}$ is a categorical variable, so that $\boldsymbol{y} \in \mathbb{Z}^n$. It is important to note that for classification tasks, the class labels must be known to the model, meaning that they must appear in the training data [102, p 3]. A task in which samples can only belong to one of two groups is called a binary classification problem. A model trained to predict class labels of samples is called a classifier. As experiments carried out in the course of this work only rely on support vector machines (SVMs) and random forests (RFs), only these two models are covered in the following sections.

3.1.2 Support Vector Machine

An SVM is an algorithm that constructs linear decision boundaries, also called hyperplanes, to separate the feature space. As it always finds the largest margin that separates two classes, it is called a maximum margin classifier. Originally, an SVM could only be used for binary classification cases. However, two common approaches, One-Against-One (1A1) and One-Against-All (1AA), exist, to overcome this limitation [7, p 2]. In 1AA, the n_c-class problem, where n_c denotes the number of classes subject to the classification problem, is transformed into n_c 2-class problems by comparing each class to all other observations that are not in this particular class [7, p 2]. 1A1, on the other hand, constructs one machine for each pair of classes, which results in $n_c(n_c - 1)/2$ hyperplanes [7, p 2]. However, the binary classification case is the starting point for both of these approaches.

Separable Case

In the separable case, a hyperplane can be defined so that all observations of both classes are completely separated by that hyperplane. The decision is then made for a new sample $\boldsymbol{u}$ based on its position with regard to that hyperplane. The function for the hyperplane is $h(x) = b + \boldsymbol{w}^T\boldsymbol{x}$ with $\boldsymbol{w}$ being the weights and b being the bias [11, p 43]. At training stage, the SVM learns the weights $\boldsymbol{w} = [w_1, w_2, \dots w_p]^T$ and the bias b, which is a scalar, so that the distance to the hyperplane is at least 1 for each data point [11, p 44]. As the SVM finds the maximum margins, $h(x)$ will be 1 for only the nearest data points to the margin for each class, the so-called support vectors.

For the sake of symmetry, binary class membership can be represented by $\boldsymbol{y} \in \{-1, 1\}^n$ [156, p 11]. By using this variable, the constraint

$$y_i(\boldsymbol{x}_i\boldsymbol{w} + b) - 1 = 0 \tag{3.1}$$

can be defined for support vectors. This constraint can then be used to compute the margin width, as the margin width is given by the dot product of a unit vector normal to the hyperplane and the difference vector of the two sides of the margin, which is defined by the support vectors. This means that, using h_y as the margin border for the respective class labels $y_i \in \{-1, 1\}$, the margin $M = (\boldsymbol{h}_1 - \boldsymbol{h}_{-1}) \cdot \frac{\boldsymbol{w}}{||\boldsymbol{w}||}$. The margin width is given as

$$M = \frac{1 - b - (-1 - b)}{||\boldsymbol{w}||} = \frac{2}{||\boldsymbol{w}||} \tag{3.2}$$

using the constraint (3.1). Maximizing the margin is achieved by either maximizing $\frac{1}{||\boldsymbol{w}||}$ or by minimizing $||\boldsymbol{w}||$, which means that the learning problem can be defined as

$$\min_{\boldsymbol{w},b} \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} \tag{3.3}$$

for mathematical convenience [156, p 14]. To find the extremum of the function, Lagrange multipliers can be used [11, p 44]. This leads to the Lagrange primal function

$$L_p = \frac{1}{2}||\boldsymbol{w}||^2 - \sum_{i=1}^{n} \alpha_i [y_i(\boldsymbol{x}_i\boldsymbol{w} + b) - 1] \tag{3.4}$$

where the Lagrange multipliers will be non-zero only for support vectors. Then, the partial derivatives need to be taken in respect to $\boldsymbol{w}$ and b and be set to 0.

$$\frac{\partial L_p}{\partial \boldsymbol{w}} = \boldsymbol{w} - \sum_{i=1}^{n} \alpha_i \boldsymbol{x}_i y_i = 0 \Rightarrow \sum_{i=1}^{n} \alpha_i \boldsymbol{x}_i y_i = \boldsymbol{w} \tag{3.5}$$

$$\frac{\partial L_p}{\partial b} = \sum_{i=1}^{n} \alpha_i y_i = 0 \tag{3.6}$$

These derivatives lead to the Lagrange dual function

$$L_d = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j y_i y_j \boldsymbol{x}_i^T \boldsymbol{x}_j \tag{3.7}$$

which denotes the lower bound of the objective function [11, p 44]. The optimization depends only on the dot product of pairs of samples. The final optimization problem is

$$\max_{a_i} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j y_i y_j \boldsymbol{x}_i^T \boldsymbol{x}_j. \tag{3.8}$$

Non-Separable Case

In the non-separable case the two classes overlap and can, therefore, not be fully separated. To address this problem, a slack variable $\xi_i \geq 0$ is added to each side condition. ξ_i holds the magnitude of the violation for each sample, which is given by the amount of which the sample is on the wrong side of the margin. If a data point is correctly classified and lays outside the margin, the slack variable $\xi_i = 0$. This new variable changes the side condition (3.1) to

$$y_i(\boldsymbol{x}_i\boldsymbol{w} + b) \geq M(1 - \xi_i) \tag{3.9}$$

which sets the violation in relation to the margin. Furthermore, the minimization problem (3.3) is changed to

$$\min_{\boldsymbol{w},b} \left[\frac{1}{2}||\boldsymbol{w}||^2 + C\sum_{i=1}^{n} \xi_i \right] \tag{3.10}$$

where C is the cost parameter, which is used to tune the margin. The harder $\sum_{i=1}^{n} \xi_i$ is penalized, the harder the margin becomes [11, p 47]. The optimization problem is the same as in equation (3.8), except that $0 \leq a_i \leq C$ for $i = 1, \ldots, n$.

Kernel Trick

By design, an SVM only supports linear decision boundaries. However, the kernel trick can be employed to elevate the features to a high dimensional kernel space $\mathbb{R}^\kappa$ with $p < \kappa$ [85, p 5]. The kernel method is the transformation from feature space to kernel space and is defined as $\phi(\cdot) : \mathbb{R}^p \to \mathbb{R}^\kappa$ [84, p 167]. By employing the transformation in the kernel function $k(\cdot, \cdot) : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$ the optimization problem can be solved without knowing the explicit mapping to the kernel space as the function

$$K(\boldsymbol{u}, \boldsymbol{v}) = \langle \phi(\boldsymbol{u}) \ , \ \phi(\boldsymbol{v}) \rangle \tag{3.11}$$

computes the inner product in the kernel space [84, p 167]. When employing the kernel trick, the Lagrangian dual function (3.7) becomes

$$L_d = \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j y_i y_j K(\boldsymbol{x}_i, \boldsymbol{x}_j). \tag{3.12}$$

Common kernel functions implemented in popular machine learning libraries like scikit-learn [136] are the following:

1. Linear kernel: $K(\boldsymbol{u}, \boldsymbol{v}) = \langle \boldsymbol{u} \ , \ \boldsymbol{v} \rangle = \boldsymbol{u}^T\boldsymbol{v}$.

2. Polynomial kernel: $K(\boldsymbol{u}, \boldsymbol{v}) = (\gamma\langle \boldsymbol{u}\ ,\ \boldsymbol{v}\rangle + r)^d$, where r is a constant, $\gamma > 0$, and d is the degree of the polynomial.
3. Radial basis function kernel (RBF): $K(\boldsymbol{u}, \boldsymbol{v}) = exp(-\gamma||\boldsymbol{u} - \boldsymbol{v}||^2)$ with $\gamma > 0$.
4. Sigmoid kernel: $K(\boldsymbol{u}, \boldsymbol{v}) = tanh(\gamma\langle \boldsymbol{u}\ ,\ \boldsymbol{v} + r)$, where r is a constant and $\gamma > 0$.

3.1.3 Random Forest

In contrast to SVMs, RFs are ensembles of individual classifiers called decision trees.

Decision Tree

Decision trees are directed acyclic graphs with only one root node and at most one path between every pair of nodes [71, p 2]. Each node contains a decision rule based on certain features. For univariate decision trees, each node is associated with one feature $\rho \in \boldsymbol{\mathcal{F}}^p$ and each edge from outgoing from that node is associated with one or more values of that feature [71, p 3]. However, a decision rule could consider multiple features as well.

Let Q be the data in node Q_m. Furthermore, let j_ρ be the index of feature $\rho \in \boldsymbol{\mathcal{F}}^p$ in $\boldsymbol{X}$. As far as Q_m is not a leaf node, Q_m has two outgoing edges to Q_l and Q_r that are related to the a splitting rule $\Phi = (\rho, \boldsymbol{X}_{:,j_\rho})$, where ρ is the feature and $\boldsymbol{X}_{:,j_\rho}$ is a collection of potential splitting points. The loss function for a node can then be defined as

$$L(Q, \Phi) = \frac{n_l}{n_m} H(Q_l) + \frac{n_r}{n_m} H(Q_r), \tag{3.13}$$

where n_l, n_m, and n_r are the number of samples falling in the respective nodes Q, Q_l, and Q_r and $H(\cdot)$ is an impurity function [169, p 3]. A decision tree is then built recursively by computing the optimal split $\hat{\Phi}_m$ for node Q_m

$$\hat{\Phi}_m = \min_{\Phi} L(Q_m, \Phi) \tag{3.14}$$

that minimizes the loss function [169, p 3]. According to [169, p 3], misclassification error H_M, gini index H_G, and cross-entropy H_C are common choices for the impurity function H, and they are defined as

$$H_M(m) = 1 - \max_{1 \leq k \leq K} p_{mk} \tag{3.15}$$

$$H_G(m) = 1 - \sum_{k=1}^{K} p_{mk}^2 \tag{3.16}$$

$$H_C(m) = -\sum_{k=1}^{K} p_{mk}\ log\ p_{mk}, \tag{3.17}$$

where K are the class labels and

$$p_{mk} = \frac{1}{n_m} \sum_{\boldsymbol{x}_i \in m} \mathbb{1}(y_i = k). \tag{3.18}$$

This way, decision trees with optimal splitting conditions can be constructed. However, this also means that fully grown decision trees are highly sensitive to the training data, which leads to a high risk of overfitting to the training data and low generalization. Pruning, which is the action of removing unreliable branches of the decision tree, can be used to lower these risks [109, p 228]. Another way to solve this problem is by using ensemble methods like RFs.

Ensemble Methods

RFs were introduced to preserve the advantages of decision trees while preventing the classifier to overfit [65, p 1]. RFs create B random decision trees $\boldsymbol{T} = [T_1, T_2, \ldots, T_B]$ by randomly sampling the training data with replacement and considering only a subset of features $\boldsymbol{\mathcal{F}}_b$ for each decision tree T_b [66, p 834]. Since each tree is trained using a different proper subset of features $\boldsymbol{\mathcal{F}}_b \subset \boldsymbol{\mathcal{F}}$, each tree generalizes for different unselected dimensions [66, p 833]. According to [35, p 2], an RF classifier $\hat{C}_{rf}^B$ then predicts the class label by majority vote

$$\hat{C}_{rf}^B(\boldsymbol{x}_i) = \texttt{majority}\left\{T_b(\boldsymbol{x}_i)\right\}_1^B = \hat{y}_i. \tag{3.19}$$

Another advantage of RFs is that the feature importance can be computed from the final model. The feature importance takes into account the impurity decrease when reaching a node Q_m weighted by the chance of reaching that node. The chance of reaching node Q_m can be defined as $w_m = \frac{n_m}{n}$. A node's importance ι_m is then given as

$$\iota_m = w_m H(Q_m) - w_l H(Q_l) - w_r H(Q_r). \tag{3.20}$$

The feature importance τ_ρ can then be computed by

$$\tau_\rho = \frac{\sum_{q=1}^{Q^\rho} \iota_q}{\sum_{q=1}^{Q} \iota_q}, \tag{3.21}$$

where Q^ρ denotes the subset $Q^\rho \subseteq Q$ that has a splitting rule concerning feature ρ.

3.1.4 Data Splitting

As already stated in the introduction to the machine learning chapter above, a machine learning model is trained only on a subset of $\boldsymbol{\mathcal{D}}$. The basic approach called hold-out

method is to split the data $\boldsymbol{\mathcal{D}}$ into two parts $\boldsymbol{\mathcal{D}}_{train}$ and $\boldsymbol{\mathcal{D}}_{test}$ and evaluate the model fit to $\boldsymbol{\mathcal{D}}_{train}$ on the data $\boldsymbol{\mathcal{D}}_{test}$ [10, p 13]. However, this makes the performance of the model strongly dependent on the chosen split. In order to generalize this validation technique, cross-validation (CV) can be employed. This way, multiple hold-out estimators are averaged across different data splits [10, p 14]. CV can be performed in different ways. k-fold CV refers to the procedure of splitting the data into k subsets of approximately same size and then always using just one of these subsets as the validation set and the rest for training. There are also methods like Leave-one-out-CV (LOO) and Leave-p-out-CV where only p samples ($p = 1$ for LOO) are used for validation and the rest is used for training. However, these two methods are computationally very expensive.

3.1.5 Performance Evaluation Metrics

Evaluation metrics are used to measure the prediction performance of a machine learning model. There are many different evaluation metrics but this chapter will only consider metrics for binary classification.

In binary classification tasks, the model's predictions can be either correct or incorrect for each of both labels. This leads to four possible outcomes per sample:

1. True positive (TP): a positive sample was correctly classified as positive.
2. False positive (FP): a negative sample was incorrectly classified as positive.
3. True negative (TN): a negative sample was correctly classified as negative.
4. False negative (FN): a positive sample was incorrectly classified as negative.

Prediction results can then be represented in a confusion matrix

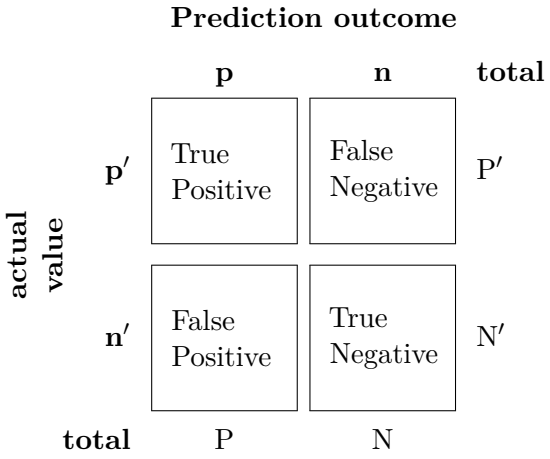


Table 3.1: Structure of a Confusion Matrix (Table taken from [104])

Common performance metrics can then be computed from the confusion matrix [50, p 862].

1. Precision: $PR = \frac{TP}{TP+FP}$.
2. Recall: $R = \frac{TP}{FN+TP}$.
3. False positive rate: $FPR = \frac{FP}{FP+TN}$.
4. Accuracy: $ACC = \frac{TP+TN}{FP+FN+TP+TN}$.
5. F1-Score: $F_1 = 2 \cdot \frac{PR \cdot R}{PR+R}$.

Graphically, a classifier's performance can be visualized in ROC space which is a two-dimensional graph, plotting the FPR on the x-axis and R on the y-axis [50, p 862]. Discrete classifiers are then represented as a dot in this graph at point (FPR , R). When drawing a line from (0 , 0) to (FPR , R) and from (FPR , R) to (1 , 1), that plot is called a ROC curve. Using this curve, another metric, the ROC AUC score, can be computed, which is the area under the ROC curve. Usually, a diagonal line from (0 , 0) to (1 , 1) is added to the plot which denotes a classifier's performance that makes random predictions.

3.1.6 Hyperparameter Tuning

Hyperparameter tuning is the process of repeatedly training and evaluating a machine learning model with different parameters $\boldsymbol{\theta}_i$ on the same data. This way, the model is tuned to the data. However, this process is not done by the model itself, but it is typically a manual process using a predefined hyperparameter grid with G different combinations $\boldsymbol{\Theta} = [\boldsymbol{\theta}_0, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_G]$ [13, p 199]. By evaluating the model for each combination $\boldsymbol{\theta}_i$ in $\boldsymbol{\Theta}$, the best set of parameters $\boldsymbol{\theta}^* \in \boldsymbol{\Theta}$ can be found. Parameter configurations are usually compared using CV.

3.1.7 Data Preprocessing

Data $\boldsymbol{\mathcal{D}}$ can often not be directly used to build a model due to missing values, inappropriate data types, or data properties that would distort the model. However, it is important to preprocess the data after splitting to assure independence of training and test data.

Missing Data

A sample $\boldsymbol{x}_i$ is considered to have missing values if at least one of its features does not hold a value. Since many models require each sample to have values for each feature, samples with missing values need to be processed first. Generally, there are two options: the sample can be discarded or the respective value can be imputed.

Deleting samples with missing values reduces the sample size and is therefore often not practicable. Data imputation, on the other hand, introduces noise to the samples. Values can be imputed in many ways like:

1. Mean imputation: Take the mean of the training set.
2. Median imputation: Take the median of the training set.
3. Hot deck imputation: Take the value of a random instance of similar samples.
4. Cold deck imputation: Take the value of a specific instance of similar samples.

The right imputation technique depends on the data. In order to find the best technique, different imputation techniques can be tested as part of Θ in a grid search.

Data Types

In many cases, $\boldsymbol{\mathcal{D}}$ does not only contain continuous numeric data but also categorical data encoded as strings. In these cases, the data needs to be encoded for the model to be processable. However, simply encoding categories as numbers might induce unwanted implications as suddenly an order between categories

$$(\texttt{red}\ ,\ \texttt{blue}) \xrightarrow{\text{encoding}} (1\ ,\ 2) \Rightarrow \texttt{red} < \texttt{blue} \tag{3.22}$$

is established.

For this reason, one-hot-encoding is widely used to encode categorical data. Let v_ρ be the distinct values of ρ in $\boldsymbol{\mathcal{D}}$ and $|v_\rho|$ be the number of distinct values. One-hot-encoding then creates $|v_\rho| - 1$ new binary variables. Each column then indicates the presence or absence of one categorical value. Thus, this encoding potentially increases the data's dimensionality.

Feature Scaling

Since some machine learning models use distance measures between data points, a feature with higher values potentially influences the distance more than a feature with lower values. For this reason, feature scaling is important, as it balances the impact of features on the distance measure. However, there are different methods for scaling features available, and the choice can be made by testing their impact on the performance as part of Θ in a grid search.

Common scaling methods are:

1. Standardization: $x'_{i,\rho} = \frac{\boldsymbol{x}_{i,\rho} - \mu_\rho}{\sigma_\rho}$ which leads to $\mu'_\rho = 0$ and $\sigma'_\rho = 1$.
2. Min-Max-Scaling: $x'_{i,\rho} = \frac{\boldsymbol{x}_{i,\rho} - \min(\boldsymbol{x}_{:,\rho})}{\max(\boldsymbol{x}_{:,\rho}) - \min(\boldsymbol{x}_{:,\rho})}$ which leads to $0 \leq x'_{i,\rho} \leq 1$.

3.2 Similarity

Similarity is a value between 0 and 1 which is computed by a similarity function $s(\cdot\ ,\ \cdot)$. This function can also be expressed as the inverse of a corresponding distance function $d^{-1}(\cdot\ ,\ \cdot)$. Let o and u be objects of any type. Then these objects can be similar in many ways. For example, let s_o be the name of o, s_u be the name of u, m_o be the meaning of s_o and m_u be the meaning of s_u. Then, $s(s_o\ ,\ s_u)$ could result in a high similarity while $s(m_o\ ,\ m_u)$ finds only a low similarity. This means that similarity always refers to specific aspects of two objects. Methods to compare these different aspects relevant to trademark data are described in the following sections.

3.2.1 Semantic Similarity

Semantic similarity refers to the relatedness of the semantic meaning of two strings. This means that, in contrast to string similarity discussed in section 3.2.2, strings are not compared lexicographically. However, this also means that a function that makes semantic comparisons is required to have knowledge about the semantic meaning of strings. This is done with so-called word embeddings which are vector representations of words incorporating their semantic and syntactic meanings [155, p 1].

There are multiple methods that can be used to create word embeddings. Word embeddings are learnt in a specific context which is defined by the corpus used to trained the model. This limits each trained model to a vocabulary W used in the corpus. However, hybrid models that compare unknown words on character level can be used to overcome this problem and allow for open vocabulary embeddings [101, p 1].

Continuous-Bag-of-Words (CBOW) and skip-gram are two iteration-based methods proposed in [108]. These methods have a similar architecture but make predictions in different ways.

CBOW predicts the word in the center of a so-called window which is another word for the surrounding words within a predefined range, the so-called window size [155, p 2]. This means, that it maximizes the probability of a word being in a specific context

$$P(w_i|w_{i-c}, w_{i-c+1}, \ldots, w_{i-1}, w_{i+1}, \ldots, w_{i+c-1}, w_{i+c}), \tag{3.23}$$

where c is the window size and w_i is the word at position i [155, p 2]. Therefore, the CBOW model trains two matrices, an input word matrix $\boldsymbol{V} \in \mathbb{R}^{N\times|W|}$, where each column in $\boldsymbol{V}$ corresponds to a word in the vocabulary, and the values in the column are N-dimensional embedded vectors representing those words, and an output matrix $\boldsymbol{U} \in \mathbb{R}^{|W|\times N}$, where each row represents a word in $\boldsymbol{W}$, and the values in the row are N-dimensional embedded vectors for those words [155, p 2]. The model starts with a one-hot representation for each word which is then multiplied by $\boldsymbol{V}^T$ to obtain N-dimensional word vector embeddings [155, p 2]. A probability vector for an input word can be obtained by applying $\boldsymbol{U}^T$ to the input word and then employing the softmax

operation [155, p 2]. Then, minimizing the cross-entropy loss between the probability vector and the embedded vector of the output word yields the CBOW model [155, p 2].

The skip-gram model, on the other hand, maximizes the probability of the context given a word

$$P(w_{i-c}, w_{i-c+1}, \ldots, w_{i-1}, w_{i+1}, \ldots, w_{i+c-1}, w_{i+c}|w_i). \tag{3.24}$$

Other than that, both models work similarly [155, p 2]. However, these models are limited to information in their local context windows and cannot leverage statistics of the whole corpus [121, p 1532]. The Global Vectors (GloVe) method adjusts the skip-gram architecture by considering also the co-occurences of words, meaning how often any word w_j occurs in the context of a word w_i [121, p 1533]. The GloVe methods outperformed CBOW and skip-gram models in word analogy, word similarity, and named entity recognition tasks [121, p 1541].

The similarity between to words is given by the cosine similarity

$$cos(\boldsymbol{w}_x, \boldsymbol{w}_y) = \frac{\boldsymbol{w}_x \cdot \boldsymbol{w}_y}{||\boldsymbol{w}_x||\ ||\boldsymbol{w}_y||}, \tag{3.25}$$

where $\boldsymbol{w}_x$ and $\boldsymbol{w}_y$ are word vectors of the respective words to be compared and $||\boldsymbol{w}_x||$ and $||\boldsymbol{w}_y||$ are their ℓ_2 norm [155, p 4]. The similarity of words represented as nodes c_1 and c_2 in an ontology can be computed using the Wu-Palmer similarity

$$sim_{wp}(c_1, c_2) = \frac{2 \cdot depth(lcs(c_1, c_2))}{depth(c_1) + depth(c_2)}, \tag{3.26}$$

where lcs is the lowest common subsumer in the ontology and depth(c) denotes the depth of node c [163, p 136].

3.2.2 String Similarity

In this work, string similarity refers to syntactic similarity that does not consider anything but the spelling of a word. Thus, the semantic meaning of words is not taken into account. There are two categories of syntactic similarity measures, namely character-level and token-level measures [55, p 170].

Character-level measures

Character-level measures view strings as a sequence of characters. Sequences of characters s_1 and s_2 can be compared either by the number of edit operations required so that s_1 equals s_2 or by the longest common substring (LCS) [55, p 171].

Table 3.2 is taken from [55, p 172]. It demonstrates the differences of popular similarity measures, where s_1 and s_2 are the input strings, v denotes variable cost, m is the number

Similarity measure	Equation	Operation cost			
		Ins	Del	Sub	Swap
Levenshtein [97]	$1 - \frac{edit(s_1\ ,\ s_2)}{max(\lvert s_1\rvert,\lvert s_2\rvert)}$	1	1	1	-
Damerau-Levenshtein [31]	$1 - \frac{edit(s_1\ ,\ s_2)}{max(\lvert s_1\rvert,\lvert s_2\rvert)}$	1	1	1	1
Needleman & Wunsch [114]	$1 - \frac{edit(s_1\ ,\ s_2)}{2\times max(\lvert s_1\rvert,\lvert s_2\rvert)}$	v	v	1	-
Smith & Waterman [144]	$\frac{edit(s_1\ ,\ s_2)}{min(\lvert s_1\rvert,\lvert s_2\rvert)}$	v	v	-2	-
Gotoh [59]	$\frac{edit(s_1\ ,\ s_2)}{min(\lvert s_1\rvert,\lvert s_2\rvert)}$	v	v	± 3	-
Hamming [64]	$1 - \frac{edit(s_1\ ,\ s_2)}{max(\lvert s_1\rvert,\lvert s_2\rvert)}$	-	-	1	-
Jaro [74]	$\frac{1}{3} \times \left(\frac{m}{\lvert s_1\rvert} + \frac{m}{s_2} + \frac{m-x}{m}\right)$	-	-	-	-
Jaro-Winkler [160]	$J(s_1\ ,\ s_2) + (l \times p(1 - J(s_1\ ,\ s_2)))$	-	-	-	-
LCS [53]	$\frac{\lvert sub(s_1\ ,\ s_2)\rvert}{max(\lvert s_1\rvert,\lvert s_2\rvert)}$	-	-	-	-

Table 3.2: Popular Similarity Measures [55, p 172]

of matching characters, x is the number of transposed characters divided by 2, p is a scaling factor and l is the length of the common prefix restricted to four characters. The abbreviations Ins, Del, and Sub stand for insert, delete and substitute. The cost for substitution in the Smith-Waterman-Gotoh algorithm depends on the characters subject to the substitution [55, p 172].

These algorithms are often used for sequence alignment in bioinformatics. However, they are also popular for detecting misspellings. Furthermore, some measures like the Levenshtein distance can be used with weighted operation costs, meaning that for example replacing character J with I is less expensive than replacing J with Q, which can be useful for Optical Character Recognition (OCR) as it also takes into account the visual similarity of characters [62, p 3].

Token-level measures

Token-level measures operate on a higher level than character-level measures. A token is a segment of a string obtained by either using the q-grams approach [140] or tokenization [55, p 172]. Q-grams create tokens by sliding a window of size q over the string. Another method is to use 1-skip-grams which build tokens from the surrounding characters. Tokenization splits a string using so-called delimiters. Delimiters are characters which separate tokens from each other. For example, words can be obtained as tokens from a sentence if whitespaces are defined as delimiters.

The similarity of tokens can then be computed by three methods: sequence matching, set

matching, and the bag-of-tokens method [55, p 173 f]. The sequence matching method works just like character-level measures by viewing each token as a unit [55, p 173]. Set matching does not consider the order of tokens but only computes the overlap between two sets [55, p 173]. However, this approach is still very dependent on the spelling of words as the strings *"gray color"* and *"grey colour"* would have a similarity of 0 [55, p 173]. This problem can be solved by combining this approach with a character-level measure so that tokens are considered matching if their character-level distance is below a certain threshold [107]. The bag-of-tokens method measures the term frequency-inverse document frequency (TF-IDF)

$$tfidf(t_i) = \frac{f_{t_i}}{F} \cdot log\left(\frac{N}{N_i}\right), \tag{3.27}$$

where t_i is a token, f_{t_i} is the frequency of t_i, F is the frequency of the most frequent token in all strings, N is the number of strings compared and N_i is the number of strings containing token t_i [3] [55, p 174].

3.2.3 Image Similarity

The similarity of images is impacted by color, spatial, and temporal properties as they lead to the perception of structures, regions, and shapes [92, p 265]. This means that a low-level pixel-wise comparison of two images will most likely not result in a satisfactory similarity score. However, there are multiple features that can be extracted from images, like color histograms or spatial patterns, that sum up information from multiple pixels across the image. However, these features are still relatively low-level and cannot pick up higher-level semantics in an image.

To overcome this problem, the fully connected layers of convolutional neural networks (CNNs) can be used as feature vectors [165, p 143]. CNNs use local, convolutional feature maps that are applied in subsequent layers to increase the complexity and level of abstraction with each layer [91, p 400]. This way, CNNs are able to learn meaningful and complex representations of images. By extracting the data from fully connected layers in CNNs, it is possible to retrieve features that can then be compared using cosine similarity (3.25) [165, p 240 f].

3.3 Phonetic Encoding

Phonetic encoding refers to the transformation of a word to a representation of its pronunciation. This method allows to assess the aural similarity of words. Phonetic encodings depend on the pronunciation of words and are therefore specific to languages.

One popular algorithm for phonetic encodings is the SoundEx algorithm [154, p 346]. This algorithm retains the first letter of the word and encodes the rest with numbers according to a dictionary for letter encodings. However, the result is restricted to a

length of four characters, meaning that long words are cut off in the encoding as can be seen in line 21 in algorithm 3.1. The current letter encoding dictionary is demonstrated in table 3.3. In line 7 of algorithm 3.1 the dictionary is used to find the suitable value for a specific character. However, not every character can be found in the table. This is because not all characters, for example character H, are relevant to the phonetic encoding in this algorithm.

Keys	**Value**
B, P	1
F, V	2
C, S, K	3
G, J	4
Q, X, Z	5
D, T	6
L	7
M, N	8
R	9

Table 3.3: Modified Version of the SoundEx Encoding Table [154, p 347]

Another algorithm, NYSIIS, results in a quite different representations. It uses more complex rules to encode the string and obtains slightly better results than SoundEx [154, p 347]. While SoundEx does not pay attention to the position of each character, NYSIIS has specific substitution rules for prefixes and suffixes as can be seen in table 3.4. It is important to note that these substitutions are executed in a specific order as can be seen in pseudocode 3.2. However, NYSIIS is also restricted to a certain output length. This potentially leads to loss of information.

Metaphone is the name of a more sophisticated algorithm that creates an output of arbitrary length just consisting of letters [154, p 349]. It adds more complexity to the previous algorithms as it consists of 16 steps and its substitution rules often target sequences of multiple characters [154, p 349]. However, this algorithm was further developed. The current version, called Metaphone 3, achieves an accuracy of 98% in the task of word identification [154, p 350].

The choice of encoding algorithm depends on the task and the data. The algorithms described above typically served the purpose of identifying names even when they were misspelled. Also, the choice of encoding algorithm strongly depends on the language of the data as encoding algorithms are tuned specific languages.

Algorithm 3.1: SoundEx

Input: String s
Output: String t

```
1  Initialize d as a dictionary containing letter encodings;
2  Initialize r = [ ];
3  Declare prev;
4  s ← upper(s);
5  for index ← 0 to length(s) − 1 do
6  |   c ← s[index];
7  |   p ← d[c];
8  |   if index = 0 then
9  |   |   add c to r;
10 |   else if p and p ≠ prev then
11 |   |   add p to r;
12 |   else if p = 0 then
13 |   |   p ← null;
14 |   else
15 |   |   p ← prev;
16 |   end
17 |   prev ← p;
18 end
19 t ← join r to string;
20 if length(t) > 4 then
21 |   t ← t[0 : 3];
22 else if length(t) < 4 then
23 |   for i ← 1 to 4 − length(t) do
24 |   |   t ← t + "0";
25 |   end
26 end
27 return t
```

Keys	**Value**	**Position**
MAC	MCC	Prefix
KN	N	Prefix
K	C	Prefix
PH, PF	FF	Prefix
SCH	SSS	Prefix
EE, IE	Y	Suffix
DT, RT, RD, NT, ND	D	Suffix
EV	AF	Any
A, E, I, O, U, W	A	Any
Q	G	Any
Z	S	Any
M, KN	N	Any
K	C	Any
SCH	SSS	Any
PH	FF	Any

Table 3.4: Encoding Table for NYSIIS [154, p 347 f]

Algorithm 3.2: NYSIIS

```
   Input: String s
   Output: String t
 1 Initialize d as a dictionary containing letter encodings;
 2 Declare t;
 3 s ← upper(s);
 4 forall (key, value) in d where d.position = prefix do
 5 |   if s starts with key then
 6 |   |   replace key with value at index 0 in s
 7 |   end
 8 end
 9 forall (key, value) in d where d.position = suffix do
10 |   if s ends with key then
11 |   |   replace key with value at index length(s) − length(key) in s
12 |   end
13 end
14 forall (key, value) in d where d.position = any do
15 |   if s contains key then
16 |   |   replace key with value in s
17 |   end
18 end
19 if s ends with "A" or "S" then
20 |   remove last character of s
21 end
22 forall vowel in s do
23 |   if subsequent character is "H" then
24 |   |   remove subsequent character from s
25 |   end
26 end
27 if s ends with "AY" then
28 |   replace suffix "AY" with "Y" in s
29 end
30 t = s;
31 if length(t) > 6 then
32 |   t ← t[0 : 5];
33 end
34 return t
```

CHAPTER 4

TMSIM-500 Dataset

In the course of this thesis, a dataset is created which contains data relevant to the assessment of likelihood of confusion and double identity. The description of this dataset is based on the structure found in existing literature, namely [21]. However, since the TMSIM-500 dataset [63] is published in the context of this thesis, the sections summary, conclusion, future work, informed consent statement, funding, institutional review board statement, acknowledgments, conflicts of interest, and author contributions are omitted.

4.1 Data Description

The TMSIM-500 dataset [63] was obtained by manually extracting information from previous EUIPO opposition decisions. The dataset consists of raw data (marks' names, images and goods and services) and likelihood of confusion factors that are similar to the ones used in [52].

(a) 003172739_contested (b) 003172739_earlier

Figure 4.1: Example of Images contained in the Dataset

The dataset consists of the variables listed in table 4.1 and a folder containing the image for each figurative mark in the dataset. Each image has a unique name based on its role

in the respective opposition case indicated by the keywords “contested” and “earlier” as can be seen in figure 4.1. Images in the dataset do not have a standardized format and were not preprocessed. In total, 500 images are contained in the dataset.

Each opposition case consists of one or more comparisons of goods and services. Each such comparison is regarded as one row in the dataset, resulting in 11815 rows in the dataset. This means that on average one case deals with 23.63 contested goods or services.

	Variable	Type	Description
1	Case ID	text	unique identifier for a case
2	Type	categorical	type of both trademarks
3	Contested Trademark	text	the contested trademark’s name
4	Earlier Trademark	text	the earlier trademark’s name
5	Visual Similarity	categorical	visual similarity of both signs
6	Aural Similarity	categorical	aural similarity of both signs
7	Conceptual Similarity	categorical	conceptual similarity of both signs
8	Degree of Attention	categorical	attention paid by target public
9	Distinctiveness	categorical	distinctiveness of the earlier mark
10	Contested Goods and Services	text	applicant’s good or service
11	Earlier Goods and Services	text	opponent’s goods and services
12	Item Similarity	categorical	similarity of goods and services
13	Opposition Outcome	categorical	outcome of the opposition case
14	Outcome	categorical	outcome for each contested item

Table 4.1: Overview of the Variables in the TMSIM-500 Dataset

4.1.1 Variables

Case ID: This variable serves as a unique identifier for a case. The Case ID is a 9-digit number with two leading zeroes, making it necessary to be stored as a string. For machine-learning tasks, the dataset should be split in a way that a case occurs in one set only so that information leakage is avoided.

Type: The type can either be “word” or “figurative”. It indicates the types of both of the marks that are subject to the respective case. In case the Type is “figurative”, corresponding images can be found in the “images” directory.

Contested Trademark: The contested trademark holds the name of the mark whose application is opposed. This field accepts any string.

Earlier Trademark: The earlier trademark holds the name of the mark on which the opposition is based. This field accepts any string.

Visual Similarity: Visual similarity describes the degree of similarity between the visual representations of both marks which has been assessed in the corresponding opposition decision. In case of two word marks this similarity is evaluated by comparing the two marks' names. For figurative marks, not the marks' names but their images are compared. This variable can hold one of eight different values which can be found in table 4.2 with their respective meaning.

Aural Similarity: Aural similarity describes the degree of similarity between the pronunciations of the two marks' names which has been assessed in the corresponding opposition decision. This variable can hold one of eight different values which can be found in table 4.2 with their respective meaning.

Conceptual Similarity: Conceptual similarity describes the degree of similarity of the perception of both marks which has been assessed in the corresponding opposition decision. This variable can hold one of eight different values which can be found in table 4.2 with their respective meaning.

Value	Description
NA	similarity not assessed
0	dissimilar or comparison impossible
1	similar to a low degree
2	similar to at least a low degree or similar to an below-average degree
3	smilar to an average degree
4	similar to at least an average degree or similar to an above-average degree
5	similar to a high degree
6	identical

Table 4.2: Similarity Scores in the TMSIM-500 Dataset

Degree of Attention: The degree of attention describes how attentive the relevant target public is when consuming goods and services that are subject to the respective opposition case. It can hold one of three different values, namely 3 (average), 4 (average to high), and 5 (high).

Distinctiveness: The distinctiveness describes the degree of distinctiveness of the earlier mark. It can hold one of four values, namely 1 (very low), 2 (low), 3 (normal), and 4 (enhanced).

Contested Goods and Services: The contested goods and services is exactly one product category for which the registration is opposed. This category is compared to all the goods and services the opposition is based on. This variable can hold any string.

Earlier Goods and Services: This variable contains the goods and services on which the opposition is based and that are the most similar to the contested goods and services.

This variable can hold any string.

Item Similarity: The item similarity is the degree of similarity that was assessed in the opposition decision between one category of contested goods and services and all of the opponent's goods and services. This variable can hold one of eight different values which can be found in table 4.2 with their respective meaning.

Opposition Outcome: This variable holds the information found in the list view of [39]. The opposition outcome can be either "upheld", "rejected", or "partially upheld". This information is important as sometimes when the marks' similarities are not sufficient for likelihood of confusion the highest degree of item similarity is assumed for reason of procedural economy.

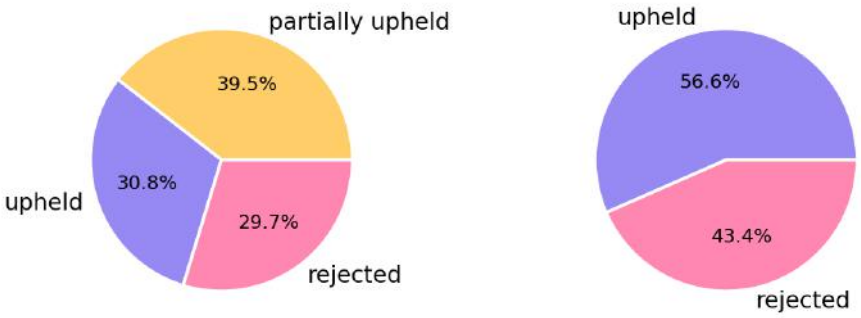
Outcome: This variable says whether the opposition is upheld or rejected for a certain contested category of goods and services. The outcome can be either "upheld" or "rejected". Predicting this variable is the goal of this thesis.

4.1.2 Value of the Data

The TMSIM-500 dataset provides information relevant to the examination of likelihood of confusion and double identity in a structured format. This allows to evaluate techniques to extract similarity scores from marks and their goods and services and to predict the outcome for each contested category of goods and services. However, this dataset does not provide information and knowledge for reasoning.

4.1.3 Data Statistics

Case Outcomes



(a) Case Outcomes (b) Outcome per Category

Figure 4.2: Proportion of Case Outcomes

Regarding the 500 cases in the dataset, the distribution of outcomes can be seen in figure 4.2(a). It can be seen that 39.5% of all cases were partially upheld. Since this dataset

is made for classification on the level of categories of goods and services, however, the outcome per category is important. These proportions are shown in figure 4.2(b).

Outcome	**Opposition-level**			**Item-level**		
	Word	**Figurative**	**Total**	**Word**	**Figurative**	**Total**
Upheld	135	98	233	4,090	2,593	6,683
Rejected	40	76	116	2,425	2,707	6,683
Partially Upheld	75	76	151	0	0	0
Total	250	250	500	6,515	5,300	13,366

Table 4.3: Outcomes per Type of Mark and Comparison Level

In table 4.3 the proportion shown in figure 4.2 are shown in absolute numbers. It becomes clear that removing the class label "partially upheld" leads to a finer granularity and creates many more samples.

Value Distribution

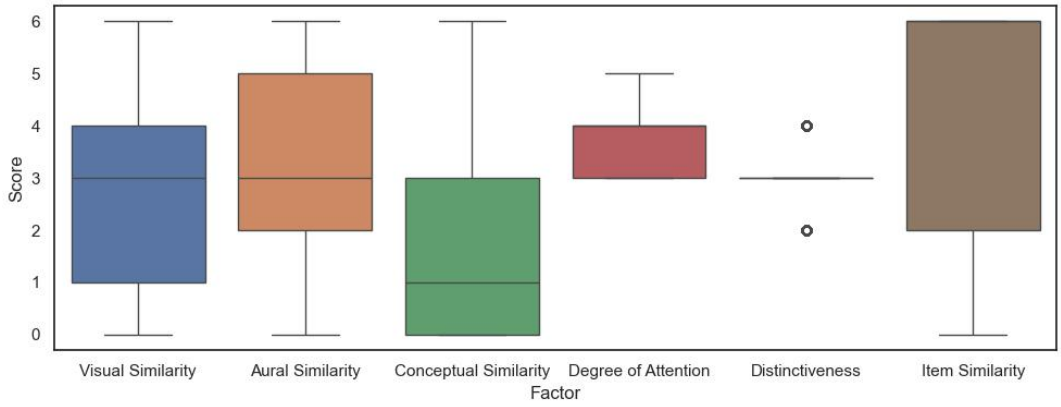


Figure 4.3: Value Distribution for Categorical Variables

The distributions of categorical variables' values are visualized in figure 4.3. It becomes apparent that only two variables, namely visual similarity and aural similarity, are almost equally distributed along the six scores. All other variables' distributions are rather skewed. Around 50% of the samples have a conceptual similarity of 0. This comes from the fact that most marks are not related to concepts which makes the conceptual comparison impossible.

The degree of attention was found to have only three possible values. Along these values, the variable is centered around 4, meaning that the target public in around 50% of the opposition cases is assumed to have an average to high degree of attention. A high degree

of attention is found in less opposition decisions than an average degree of attention. The earlier mark's distinctiveness is almost always found to be 3, meaning normal. All other values are outliers.

The item similarity is assumed to be very high in most cases. This distribution is, analogous to conceptual similarity, skewed due to the examination procedure. In cases where the signs are dissimilar, the goods and services are assumed to be identical for the reason of procedural economy.

Correlations

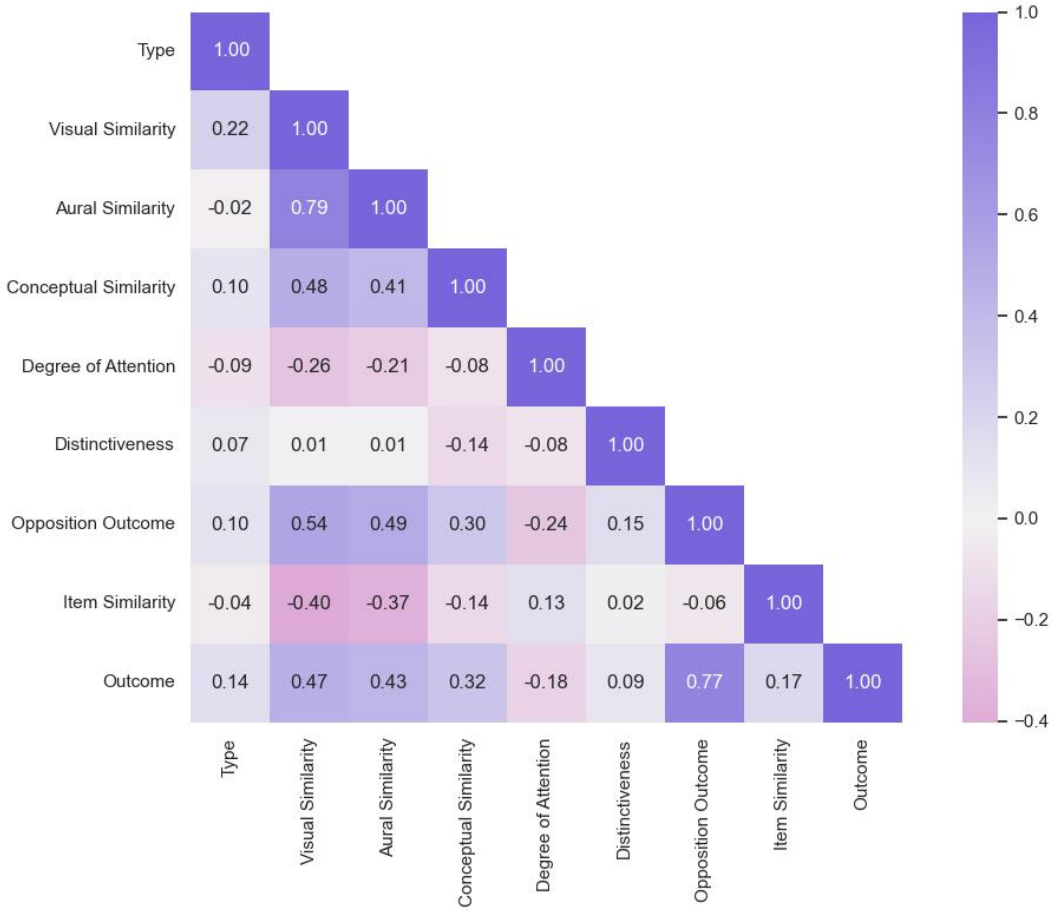


Figure 4.4: Correlation Matrix for Categorical Variables

Figure 4.4 demonstrates the correlations between the categorical variables in the dataset. It is important to note that a relatively strong correlation between visual similarity, aural similarity, and conceptual similarity is expected as they originate from the same marks. However, all other variables should not show strong correlation.

Visual similarity and aural similarity show a relatively strong correlation of 0.79. These similarity measures, however, show a rather weak correlation with the conceptual similarity. This comes from the fact that in many cases the marks do not refer to a concept which makes the comparison impossible, and thus the conceptual similarity is 0.

Another interesting observation is that visual similarity and aural similarity seem to be negatively correlated to the item similarity. The reason for this correlation is that for reason of procedural economy the goods and services are assumed to be identical. This only happens, however, when the marks are similar to a sufficiently low degree. Thus, a low similarity score lead to a high similarity score for goods and services.

Visual, aural, and conceptual similarity show a moderate correlation with the case outcome. The item similarity, however, is rather uncorrelated.

Trademark Characteristics

Trademarks' are compared by their names and, in case of figurative marks, by their figurative representation. The distribution and ordering of characters in word marks therefore has a strong impact on the visual similarity.

Statistic	**Value**
Minimum	4
Maximum	64
Mean	15.588
Median	14
0.25-Quantile	11
0.75-Quantile	18
Standard Deviation	6.960
Variance	48.445

Table 4.4: Word Mark Lengths Distribution Table

As can be seen in table 4.4 and figure 4.5, there are no word marks that consist of less than four characters. The longest word mark is 64 characters long. However, 50% of all word marks are between 11 and 18 characters long. Word marks longer than 29 characters are considered outliers. Due to the number of word marks observed, these outliers do not influence the mean a lot as the mean is quite close to the median.

Generally, there is no restriction to characters used in word marks. While the alphabet for trademarks is indefinite, the characters found in the dataset and their occurrences can be seen in figure 4.6. As it often does not matter whether characters are written in lower or upper case, figure 4.7 shows how characters are distributed without considering the character's case. However, both distributions have the same shape, meaning that the

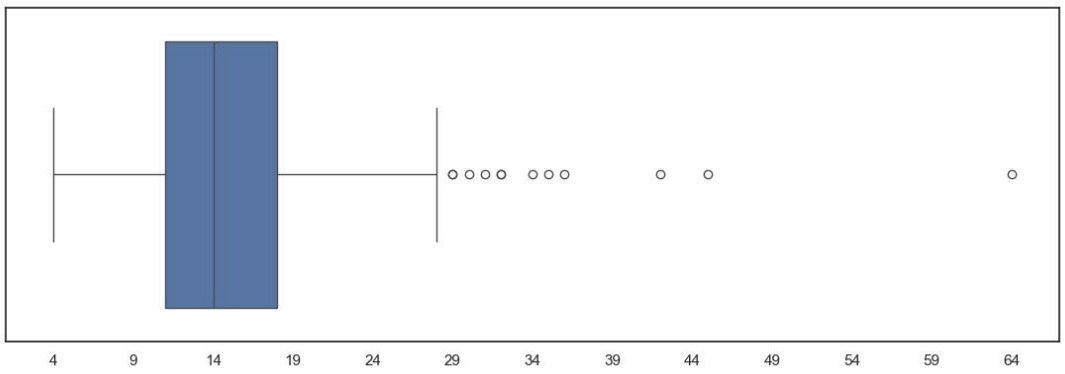


Figure 4.5: Word Mark Lengths Distribution Boxplot

uppercase vowels “E”, “A” and “O” are the most frequent characters followed by “R” and “I”. In both plots, “S”, “L” and “N” are the most frequent characters after “I”. Special characters, except for whitespaces, are rather rare. Whitespaces are more frequent than most letters, as many trademarks consists of multiple words separated by a whitespace.

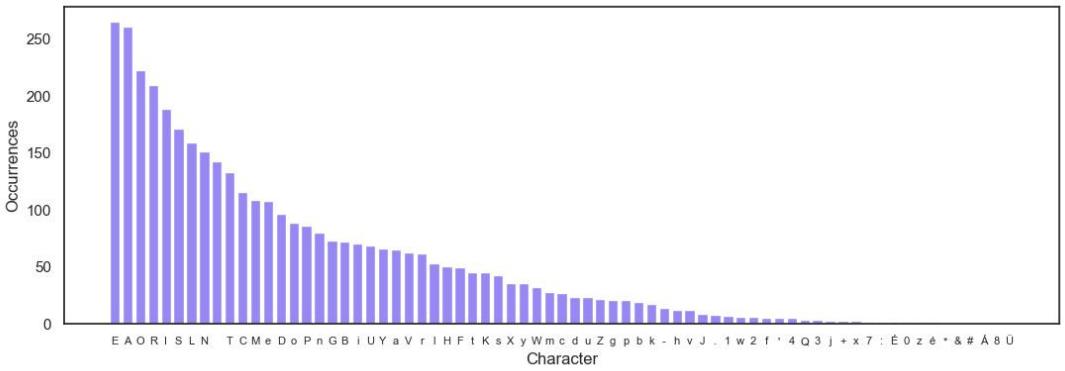


Figure 4.6: Character Frequency Distribution in Word Marks (Case sensitive)

Characters can occur in any position of a mark. There might exist a bias towards certain positions for specific characters. Figure 4.8 shows the positions for each case sensitive character. The median for most characters’ positions is between five and ten. However, there are also characters with odd distributions that can be found on the left side in figure 4.8. It is important to note that most of the characters with a skewed distribution do not occur frequently in the dataset. This means that their distribution would look different if the sample size was increased.

Apostrophes appear rather at the end of a string as many word marks have a tailing “’s”, denoting the genitive of a noun. While many characters have outliers in their distribution, these outliers come from the distribution of word mark lengths. As most word marks have

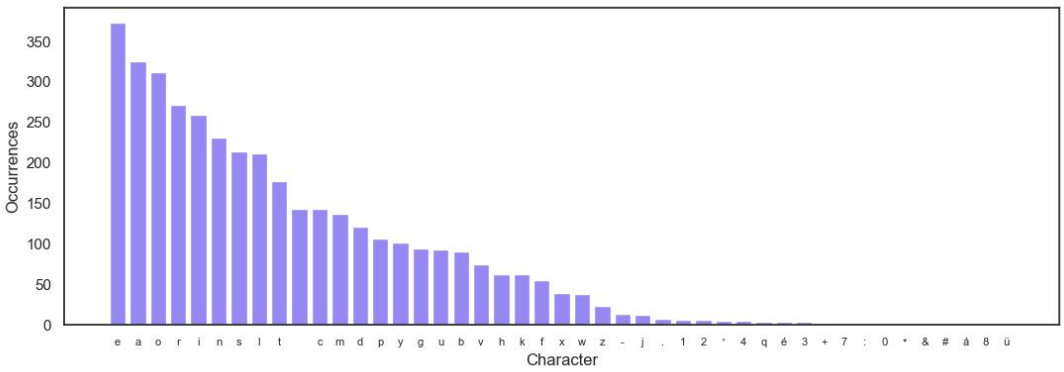


Figure 4.7: Character Frequency Distribution in Word Marks (Case insensitive)

less than 18 characters, it is obvious that there are not many instances where characters can occur at a high index.

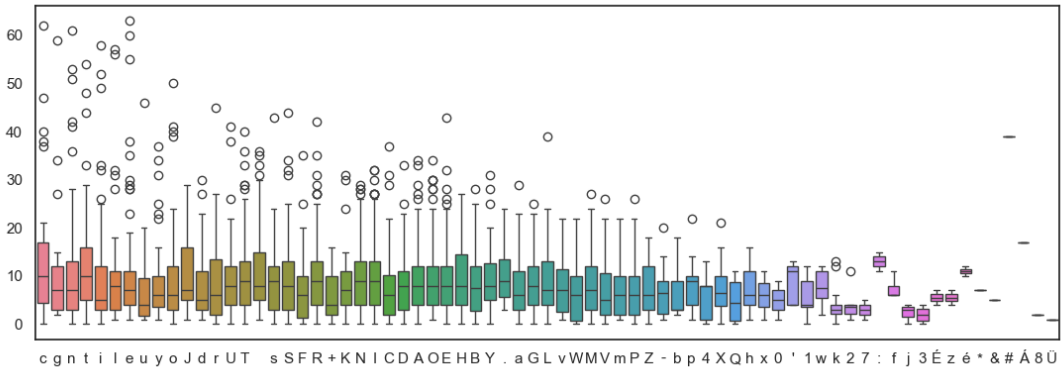


Figure 4.8: Character Position Distribution in Word Marks (Case sensitive)

Figure 4.9 shows the positions for each character without distinguishing between lower and upper case. Apart from the fact that the alphabet is smaller, there are no remarkable differences to 4.8.

Image Characteristics

TMSIM-500 contains images in the formats ".png" and ".jpg". These images can be either black and white or in color. Their size and aspect ratio, shown in figure 4.10, are arbitrary. However, all images in the dataset are wider than they are high. In figure 4.10 a dense concentration in the left bottom corner can be seen, meaning that most images appear to be less than 500 pixels wide and less than 300 pixels high.

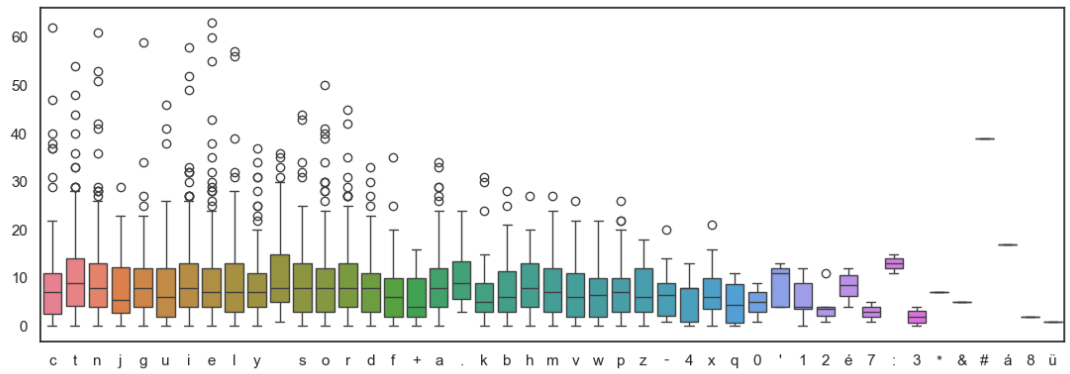


Figure 4.9: Character Position Distribution in Word Marks (Case inensitive)

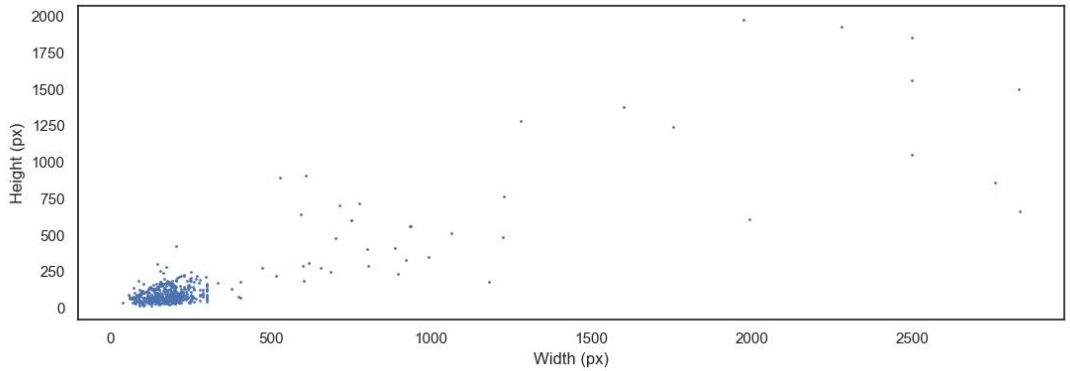


Figure 4.10: Aspect Ratios of Images in the TMSIM-500 Dataset

4.2 Method

4.2.1 Data Acquisition

Opposition decisions were taken from [39] using the search criteria listed in table 4.5.

Cases were investigated in descending temporal order. Furthermore, cases had to satisfy the following criteria.

- **Both trademarks must have a name**: This leads to a refusal of every case that contains at least one mark with figurative elements only. Marks without name are represented as "(Trade mark without text)".
- **Both trademarks must be of the same type**: Each case must either compare two figurative marks or two word marks. Regardless of the search criteria, the marks' types were double-checked on the first page of each opposition decision.

Criteria	Value
Judgment Date	Before 25/08/2023
Decision Type	Opposition Decisions
IP Right	EUTM
Language	English
Trade Mark Type	Word *or* Figurative
Opponents Earlier Right Type	Word *or* Figurative
Sort Results By	Decision/Judgment Date Descending

Table 4.5: Search Parameters used for Dataset Creation

- **The decision must be based on the assessment of likelihood of confusion or double identity**: Decisions based on the inadmissibility of an application or an insufficient proof of use are not included in this dataset. Furthermore, decisions fully based on norms other than Article 8(1) EUTMR are not considered.
- **The decision must be available in English**
- **Images must be available for each figurative mark**: In case of oppositions with two figurative marks, images must either be available in the search result list on [39] or in the opposition decision document. If no image is provided, the case must be ignored.

4.2.2 Data Labeling

The dataset was created by extracting data manually from previous opposition decisions that were selected according to the data acquisitions guidelines explained in section 4.2.1. Data points were collected per opposition decision and each variable labeled separately. In order to translate the degrees of similarity, attention or distinctiveness into categorical scores, the dictionaries in section 4.1.1 were used.

4.3 Data Availability

The TMSIM-500 dataset [63] presented in section 4 is publicly available at `https://doi.org/10.7910/DVN/PNFQLC`.

CHAPTER 5

Related Work

Likelihood of confusion and trademark law in general is a highly researched field. Research done on this topic can either focus on the legal aspects or on the computer science aspects.

5.1 Related Work in the Legal Domain

Legal research like work done by Bartow [14] and [15], Lim [99], Bone [19], Upadhye [150], Robins [131], Miaoulis and d'Amato [106], Olsen [117], Lemley and McKenna [93], Martin and Boyd [103], Rosati [133], Reinhard [130], and Coffey [29], mainly explores the factors of likelihood of confusion by analyzing case-law. These studies provide important insights into what factors played what role in specific cases. This knowledge can then be used to reason by analogy.

However, for the development of a machine learning model, a systematic empirical analysis, like work done by Beebe [16] and Blum [17], would be needed in the field of European trademark law, as the characteristics of one particular case should not influence the model. Beebe and Blum did not investigate how factors relevant to likelihood of confusion are assessed but rather in what way they contribute to the final outcome of the case.

Beebe performed their empirical study on all reported federal district court opinions from the years 2000 to 2004, resulting in a dataset of 331 opinions. This led to interesting findings, such as that, even for complex decisions, there is only a low number of decision-relevant factors, referred to as core attribute heuristic. Furthermore, decision trees based on these factors [16, p 1606] [17, p 14] and correlation matrices [16, p 1613] [17, p 18 ff] were developed. Beebe introduces a measure called multifactor stampede score, which is the difference between the proportion of factors considered that favored a finding of likelihood of confusion and the proportion of factors considered that did not favor a finding of likelihood of confusion. This measure allows to compare the influence of factors

between the different case outcomes. As a result, Beebe found that the distribution of multifactor tampede scores strongly depends on the case outcome, which can be explained by the status quo bias, meaning that the multifactor test must tilt strongly toward a likelihood of confusion in order to justify such an intervention.

Blum repeated the experiments carried out by Beebe using cases from the years 1994 to 2008. However, many opinions were removed from the dataset to assure accurate information, leading to final dataset of 206 opinions. The findings were mostly in line with the original results obtained by [16]. Both research papers support that trademark similarity is the most important factor in the multifactor test. However, Blum could not find a difference in multifactor stampede scores for cases with different outcomes in their data.

Since these findings are specific to law of the United States, they may not influence the development of TrademarkML. However, TrademarkML will compute similar statistics for data based on European trademark law.

In his book, Meitinger [105] explains how trademark applicants can search for similar trademarks by using search engines [105, p 61 ff]. Still, these approaches require legal knowledge to interpret the search results. For example, using the search engine "TMview" [46], similar images can be queried in the database. however, the most similar trademark might still not be similar enough to create a likelihood of confusion. Also, even if stated otherwise in [105, p 64], the search for marks does not seem to be reliable. For example, fuzzy search for "Frikawelle" does not list "FRIKALET" in its results and "LEICA" is not found when searching for "Leica Geosystems AG", although both pairs of marks were subject to oppositions that were upheld due to likelihood of confusion.[17] Thus, conventional methods for searching similar marks are not sufficient to reliably find conflicting trademarks.

5.2 Related Work in the Domain of Computer Science

In the field of computer science, many different problems regarding likelihood of confusion and trademark similarity have been addressed by computer scientists. The protocol for the thorough search performed to find related literature that addresses the problem of computing trademark similarities can be found in the appendix 9.

The comparison of methods, however, is difficult, since most methods are evaluated on different data. Furthermore, the data used in research is often not published nor referenced.

Most of the research on this topic addresses the problem as an image retrieval task as can be seen in tables A2, A3, and A4. While this task is quite different to the case classification task subject to this work, it also includes computing similarities between

[17] Opposition Decision of 12 June 2023, *SFK Food A/S* v *European Convenience Food GmbH*, B 003142616 and Opposition Decision of 27 June 2023, *Leica Geosystems AG* v *Tang, Qi*, B 003167556.

trademarks. This means that methods for computing trademark similarities can be employed in both tasks.

The task of content-based image retrieval (CBIR) has a rich, longstanding history in research. It is about finding relevant items in a database based on intrinsic features of items, like image similarity or semantic meaning of text. These aspects are also relevant for the assessment of likelihood of confusion. Therefore, this branch of research is included in the literature review.

In the context of CBIR, image similarity is subject to many research papers. However, the methods to address this task have changed over time. Early research investigated the performance of low-level and mid-level features that do not carry much semantic information.

Vailaya et al. [152] proposed to extract edge directions obtained by using the Canny edge dector and invariant moments, also called Hu moments, from images and then employ parametric transformations so that the input is transformed to match the trademark it is compared to. By optimizing these parameters, an energy measure is derived, which then denotes the quality of match. Follow-up research by Ciocca and Schettini [28] modified this approach by adding the mean and variance of subbands, computed using multiresolution wavelet analysis, as features. This modification improved the retrieval performance. Ravela and Manmatha [129] proposed to compare feature vectors, consisting of the principal local curvatures, computed using the local derivatives from Gaussian filtered images, and the phase, using normalized cross-covariance.

Eakins et al. [33] carried out a comparative study of several common shape-based descriptors to retrieve similar figurative marks. In their experiment, whole-image matching was compared to component-based matching. Whole-image feature sets consists of either 36 ART coefficients, seven normal moment invariants, or four affine moment invariants. The component-based feature set consists of three simple descriptors, three Rosin descriptors, eight Fourier descriptors, and the ratio of the area of each component to the area of the largest component in the image. The conducted experiments demonstrate that component-based matching of trademark images using boundary-based shape measures outperforms whole-image matching using region-based measures. However, for whole-image matching, ART coefficients outperformed normal and the affine moment invariants.

In contrast to Eakins et al., Hong and Jiang [67] propose a method combining region feature extraction and contour feature extraction to compute two marks' similarities. The presented approach guarantees invariance under rotation, translation, and scaling by normalizing. This is done by normalizing the images. First, the image is turned into a binary image, meaning that its pixel values are either 0 or 1. Then, the center of the object depicted in the image, defined by the point

$$C_m = \left(\frac{max(\boldsymbol{X}) - max(\boldsymbol{X})}{2}, \frac{max(\boldsymbol{Y}) - max(\boldsymbol{Y})}{2} \right), \tag{5.1}$$

is moved to the center

$$C_i = \left(\frac{N}{2}, \frac{N}{2}\right). \tag{5.2}$$

This process is called position normalization. Then, the object is enlarged to fit the image size. Finally, the object's rotation is normalized using the image's eigenvectors which can be found using Hotelling transform. After these preprocessing steps, region features are extracted by computing the smallest bounding circle that covers the object in the image. This circle is then split into 200 equally sized regions. Each region is then assigned a value based on whether a pixel of the object lays in that region or not. Then, corners in the image are extracted with an enhanced SUSAN algorithm and turned into contour features using the corner-to-centroid triangulations. Two trademarks are considered similar if either their region features or their contour features are similar.

Since most research used binary images, shape and texture features are dominant in early research. However, especially for trademarks, colors might contain important information. Leng and Mital [94] therefore combine invariant moments with color histograms and the image-to-background area ratio. This method was evaluated on a dataset of 100 images. Random instances were then scaled, rotated, and reshaped and used as input. The accuracy was then evaluated based on the matches returned. According to Leng and Mital, the method yields an accuracy of 95%.

Zeggari et al. [168] also proposed invariant moments and color histograms. The dataset used in their study consists of 850 original logos. The sample size was artificially increased by adding distored versions of some logos. The proposed method yields an average precision of 62% and an average recall of 74%. The authors note that their method does not support global noise. However, their method is stable under local distortions.

In a more recent study, Pinjarkar et al. [125] use color histograms, color moments, and color correlograms as color features, Gabor wavelet and Haar wavelet analysis as texture features, and fourier descriptors as well as circularity features as shape features. This method yields an average precision of 82% and an average recall of 83% on a dataset consisting of 2000 logos.

Feng et al. [51] evaluate the performance of a method combining edge features, obtained by using the Canny edge descriptor, and reversal invariant SIFT features, aggregated using the Fisher Vector, on the METU v2 dataset [149]. METU v2 was published by Tursun et al. and serves as a benchmark dataset. The proposed method using SIFT achieved a normalized average rank of 0.083.

An even better performance is achieved by Perez et al. [122]. Their method uses the VGG19 model to compute feature vectors. The similarity of these vectors is then computed using cosine similarity. Multiple configurations were evaluated, however, the best normalized average rank, 0.047, was achieved using a combination of two differently

fine-tuned VGG19 models. This approach combines a model fine-tuned to measure visual similarity and another model fine-tuned to measure conceptual similarity.

Similar experiments were carried out by Trappey et al. [147]. Their experiments were evaluated on trademark infringement cases. However, the data used in the paper was not published nor referenced. Different CNN architectures, namely AlexNet, Zfnet, SNNnet, and an improved multi-layer siemese VGG16, were compared. According to Trappey et al., the approached method for measuring image similarity yields an accuracy of 100%. However, the evaluation is based on "cases related to image similarity" [147, p 11] and no definition for such cases is given.

The same study also carries out experiments on aural similarity. A thorough comparison between SoundEx, metaphone, double metaphone, and NYSIIS is carried out. The distance between phonetic encodings is measured using a combination of the weighted levenshtein distance and LCS. The results show that the double metaphone yields an accuracy between 85.9% and 90.6%.

Anuar et al. [8] developed a model for making conceptual comparisons between trademarks in order to find matching trademarks given a query to overcame the problem of synonymy in traditional keyword-based searches. The model consists of two modules, an indexing module and a retrieval model. The indexing module is an offline component that extracts relevant conceptual features from trademarks stored in a database. It tokenizes the words and uses knowledge sources to generate extract semantic information from the respective tokens. The indexing module then stores the tokens along with their semantic information. The retrieval module then handles input queries and processes them in the same way the trademarks were processed. The conceptual similarity between the query and the trademarks can then be compared and all trademarks with a conceptual similarity above a specific threshold are then returned as relevant trademarks. Two years later, Anuar et al. [9] published results from experiments performed after implementing their formerly proposed model. The new algorithm also starts by tokenizing trademarks. However, in contrast to the initial concept, it then extracts all synonyms, hypernyms, and direct hyponyms of each token using the WordNet ontology. Then, each trademark is stored with their respective features, consisting of its tokens and their synonyms. This improves performance as the distance computation must then not be performed on the whole database but only on those trademarks that share a concept. The similarity of a trademark and a query can then be measured using the WordNet ontology with the Wu and Palmer word measure. The proposed method yields an R-precision of 66%.

Trappey et al. [147] uses a word2vec model trained on data from Google News to compute the semantic similarity of two strings. However, it is rather unlikely that a trademark name is known to the model. For such cases, Trappey et al. propose to use a vector space model that takes into account the frequency of occurrence, adjacent degree, and position order of characters. This approach, however, does not measure the similarity of related concepts anymore.

Liu et al. [100] also compared methods to classify marks as similar or dissimilar. Their

experiments only consider Chinese word marks. Liu et al. propose a model combining several embeddings on character-level. Characters are embedded using a pretrained fastText model. The authors state that this embedding also contains semantic information, however, it is important to note that there is no semantic information contained in one single character in the English language. This means that the character embedding approach will not be sufficient to compare two English words' semantics. In addition to the character embedding, Liu et al. propose a phonetic information embedding which relies on Pinyin codes. This, again, is very specific to the Chinese language and cannot directly be used for languages relying on the latin alphabet. The visual embedding of a character is obtained by taking the mean of the embeddings of its radicals, which are components of chinese characters. Furthermore, the method of Liu et al. encodes the start and end position of each word as the position of a word can influence its meaning. All these informations, the character embedding, the phonetic information embedding, the visual information embedding, and the word segmentation, are computed for each character of a word. Then, the cosine similarity between every character of each word is computed. Row-wise and Column-wise max-pooling is then performed on the similarity matrix and the position of the most similar character is stored, leading to two matrices, one matrix $\boldsymbol{A}$ for the maximum value per row and one matrix $\boldsymbol{B}$ for the maximum value per column. Two CNNs, C_A and C_B, are then fed with the position and the corresponding similarity score of the respective matrix. The output is combined with a softmax activation function.

Setchi and Anuar [139] developed a method for decision support using fuzzy logic to aggregate the overall assessment. In their research, the authors combine already existing methods to compute the visual, aural, and conceptual similarity of two trademarks with the Mamdani fuzzy inference model. First, the similarity scores are fuzzified using five triangular-based membership functions. Using 125 rules in total, five output scores can be computed which are then aggregated and defuzzified. However, this method only considers the marks' similarity scores even though goods and services of the trademarks are also important according to the interdependence principle.

CHAPTER 6

TrademarkML

Trademark monitoring plays an important role for protecting trademarks as likelihood of confusion is not assessed eo ipso. Current solutions to facilitate trademark monitoring are expensive and are often still based on manual work. Therefore, this thesis introduces the concept of TrademarkML, a trademark management system to monitor existing trademarks and trademark applications and to automatically detect conflicting trademarks. The purpose of TrademarkML is to ensure legal certainty by comparing trademarks using well-defined similarity measures and using transparent classification models to make predictions.

6.1 Concept

TrademarkML is a system that holds trademark data and allows trademark owners to be notified if the conflicting marks can be found in the European Union Trade Marks Bulletin. Furthermore, it allows applicants to check if their mark conflicts with already existing trademarks.

This basic functionality requires TrademarkML to integrate trademark and application data from existing trademark databases like [40]. Then, the system processes the data to obtain an external and an internal representation of trademarks, which are both stored. The internal representation contains encodings that are computed from the trademarks, like phonetic encodings or related concepts. The internal representation allows for a better performance, since the computation of encodings is expensive.

The core of the system is the prediction module, which is triggered for each registration and for each query. For each input consisting of similarity values between two internal representations, a binary value is returned, which indicates whether there is a likelihood of confusion between two marks and their goods and services or not. The output is

computed using a machine learning model. Each conflicting mark is reported to the user who can then decide whether or not they take further actions.

In contrast to existing solutions, results obtained by TrademarkML indicate a likelihood of confusion. Also, state of the art methods to assess the similarity of two trademarks are extended by also considering the similarity of the trademark's goods and services.

6.2 Prediction Module

6.2.1 Concept

This thesis focusses on the prediction module and the internal representation of trademarks in TrademarkML. The prediction module allows to train machine learning models and evaluate their performance given a dataset of opposition decisions. The dataset used in this thesis is presented in chapter 4. It then uses methods discussed in related work presented in section 5.2 to encode the data and compute the distance between the trademarks for each opposition decision. It then iterates over combinations of these features to find the best model and feature combination. Combinations consist of at most one feature per similarity type, like visual similarity or aural similarity. This means that this process does not combine features that are supposed to measure the same aspect of similarity.

Using this module, it is assured that experiments are carried out in the same way so that they can be compared with each other. The experiment design is discussed in section 6.2.2. By training multiple models, the best models can be found for specific performance metrics, such as F1-score, recall, or precision. This allows users to choose a model depending on how costly false negatives and false positives are for them.

6.2.2 Experiment Design

The prediction module runs experiments to find the best combination of features and the best parameters for machine learning models. This section aims to document the steps taken in the experiments. The dataset used in the experiments is presented in section 4. All methods employed are listed in table 6.2. However, it is important to note that shingle-based methods, like cosine, jaccard, n-gram, and q-gram, were used with a sequence length of 2, 3, and 4 each. Furthermore, to compute aural similarity, phonetic encodings were combined with string comparison methods. The mapping of a string to a concept in the WordNet ontology is done by finding the closest concept in the ontology using Levenshtein, cosine, and LCS.

Data Splitting

Since the dataset contains word marks and figurative marks, the first step is to separate word mark samples from figurative mark samples. Both types of marks require their own models as different features are computed for each type.

Then, each of these subsets is split into two sets, a training and a test set. The training set consists of approximately 80% of the samples while the test set consists of 20%. The method for splitting the data into training and test set makes sure that samples belonging to the same opposition case are contained in the same set. This behavior guarantees that the data in test and training set are independent from each other.

Splits are created using the GroupShuffleSplit provided by sklearn with a seed of 42. Although the grouping of samples is considered when splitting the data, the training and the test set are rather balanced and do have a similar distribution of class labels as can be seen in table 6.1.

Label	**Training Set**		**Test Set**	
	Absolute	**Relative**	**Absolute**	**Relative**
Upheld	5,442	0.56	1,241	0.61
Rejected	4,353	0.44	779	0.39

Table 6.1: Class Distribution in Training and Test Set

Internal Representation

After splitting the data, trademark information is encoded for each sample and trademark.

1. **Visual Encoding:** Images of figurative marks are encoded by feeding them to CNNs and using the output of the last fully connected layer. Three different CNNs were compared, namely VGG16, VGG19, and ResNet50.
2. **Phonetic Encoding:** Trademark names are encoded using the double metaphone algorithm and a slightly modified version of the metaphone 3 algorithm.
3. **Concept Mapping:** Trademark names are mapped to a concept by finding the minimum distance to words in the WordNet ontology. The distance measures used for this mapping are the Levenshtein distance, cosine and LCS.
4. **Item Encoding:** Goods and services of each trademark are embedded using Google's universal sentence encoder and fastText.

Feature Extraction

Features are extracted by comparing two trademarks and their goods and services using different measures. 77 features are extracted from comparisons between word marks and 56 features are extracted from comparisons between figurative marks. Each feature corresponds to a similarity aspect that has to be measured for assessing likelihood of confusion. This means that there are four categories of features. Within each category, features relate to the same similarity factor.

Since all features are computed directly from the trademarks and word marks and figurative marks are subject to different models, there are no missing values after feature extraction. This means that there is no need for deletion of samples or imputation of values.

Method	**Use**	**Reference**
Levenshtein	String Comparison	Levenshtein et al. [97]
Normalized Levenshtein	String Comparison	Yujian and Bo [166]
Damerau-Levenshtein	String Comparison	Damerau [31]
Cosine	String Comparison	Trappey et al. [147, p 5]
Jaccard	String Comparison	Gali et al. [55, p 174]
N-Gram	String Comparison	Kondrak [83]
Q-Gram	String Comparison	Gali et al. [55, p 172]
Szymkiewicz-Simpson	String Comparison	Choi et al. [27, p 44]
Jaro-Winkler	String Comparison	Winkler [160]
LCS	String Comparison	Friedman and Sideli [53]
Metric LCS	String Comparison	Bakkelund [12]
Optimal String Alignment	String Comparison	Van der Loo et al. [153, p 116]
SIFT4	String Comparison	Zackwehdex [167]
Sorensen	String Comparison	Sorensen [145]
Double Metaphone	Phonetic Encoding	Philips [123]
Metaphone 3	Phonetic Encoding	Philips [124]
VGG16 + Cosine	Image Similarity	Panagiotis Kasnesis [118]
VGG19 + Cosine	Image Similarity	Panagiotis Kasnesis [118]
ResNet50 + Cosine	Image Similarity	Panagiotis Kasnesis [118]
WordNet + Wu-Palmer	Concept Similarity	Anuar et al. [8]
Sentence Encoder + Cosine	Item Similarity	Cer et al. [22]
fastText + Cosine	Item Similarity	Bojanowski et al. [18]

Table 6.2: Methods Employed for Trademark Distance Computation

Even though these methods work in different ways, there are interesting correlations between these features. For example, most features concerning the visual similarity for word marks are highly correlated as can be seen in figure 6.1. Some features are even perfectly correlated, like optimal string alignment and Damerau-Levenshtein. Negative correlations come from the fact that some features are distance measures and some features are similarity measures. The Szymkiewicz-Simpson coefficient, referred to as

overlap in the plot, is uncorrelated to the optical string alignment and sift features.

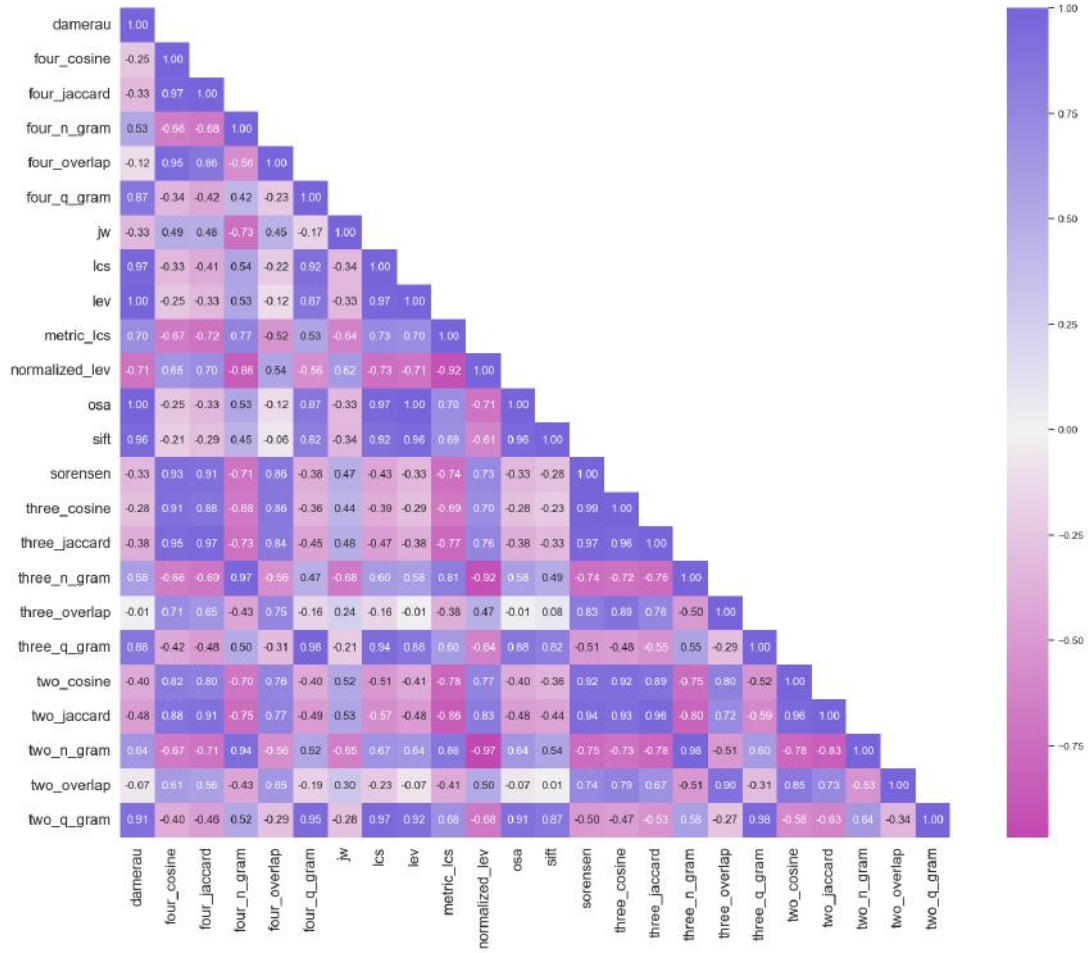


Figure 6.1: Correlation Matrix for Features concerning the Visual Similarity of Word Marks

Since the correlation between these features is high, the dimensionality can be reduced while preserving the explained variance in the lower dimensional feature space. This is done using Principal Component Analysis (PCA) after scaling the data with the RobustScaler implementation in sklearn. In fact, figure 6.2 shows that the first principal component already explains almost 93% of the variance. The first two principal components cover over 97% of the variance. However, as can be seen in figure 6.3, the selected features for determining the visual similarity of word marks alone are not sufficient to reliably distinguish between upheld and rejected oppositions. As can be seen in figure 6.5, taking into account the three first principal components does not solve this problem. A cluster of observation can still not be separated. This observation is in line with the interdependence principle mentioned in chapter 2 as visual similarity alone is not sufficient to predict

the case outcome. In figure 6.4, the first principal component of features concerning the visual similarity of word marks is plotted against the respective similarity of goods and services computed using fastText. The observations on the right side of the plot are clearly separated. However, the two variables do not contain enough information to separate the instances on the left side of the plot.

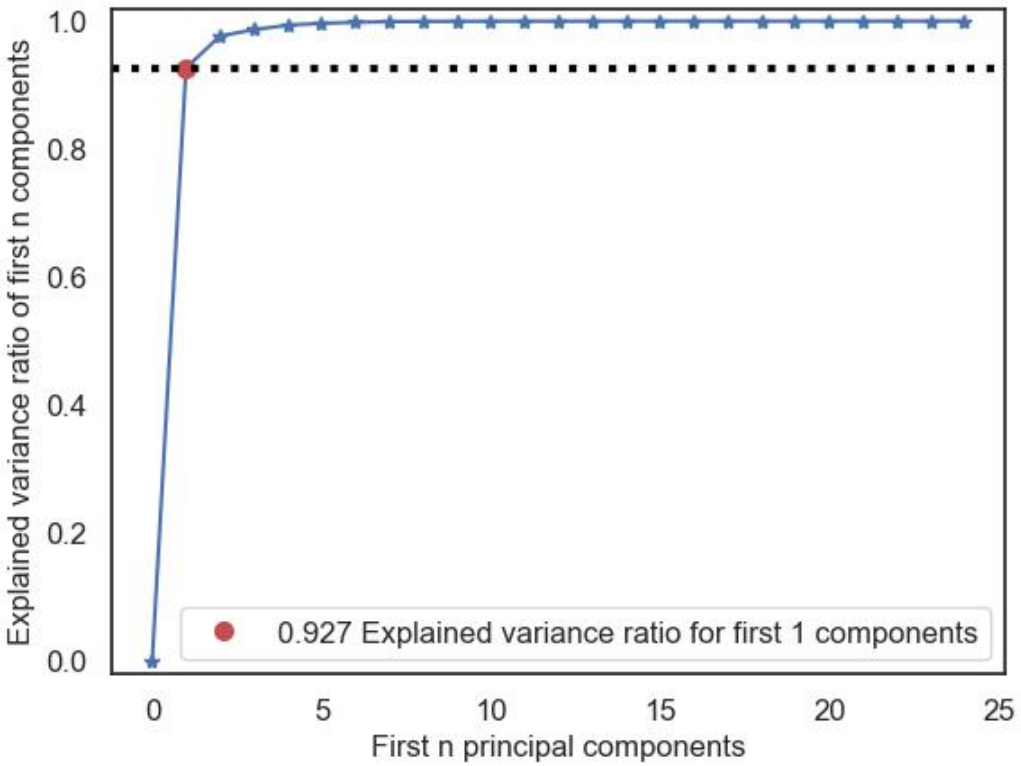


Figure 6.2: Explained Variance of the Principal Components of Features concerning Visual Similarity of Word Marks

For figurative marks, there are only three features that concern the visual similarity. Features extracted from VGG16 and VGG19 models have a high correlation of 0.9. The feature computed using the ResNet50 model, on the other hand, shares a correlation of 0.5 with the other two features, which is considered moderate. This observation is not surprising when looking at the models' architectures and the resulting activations for different layers, as can be seen in appendix A. However, even though the activations differ heavily between the VGG and the ResNet architectures, the resulting similarity scores are apparently rather correlated.

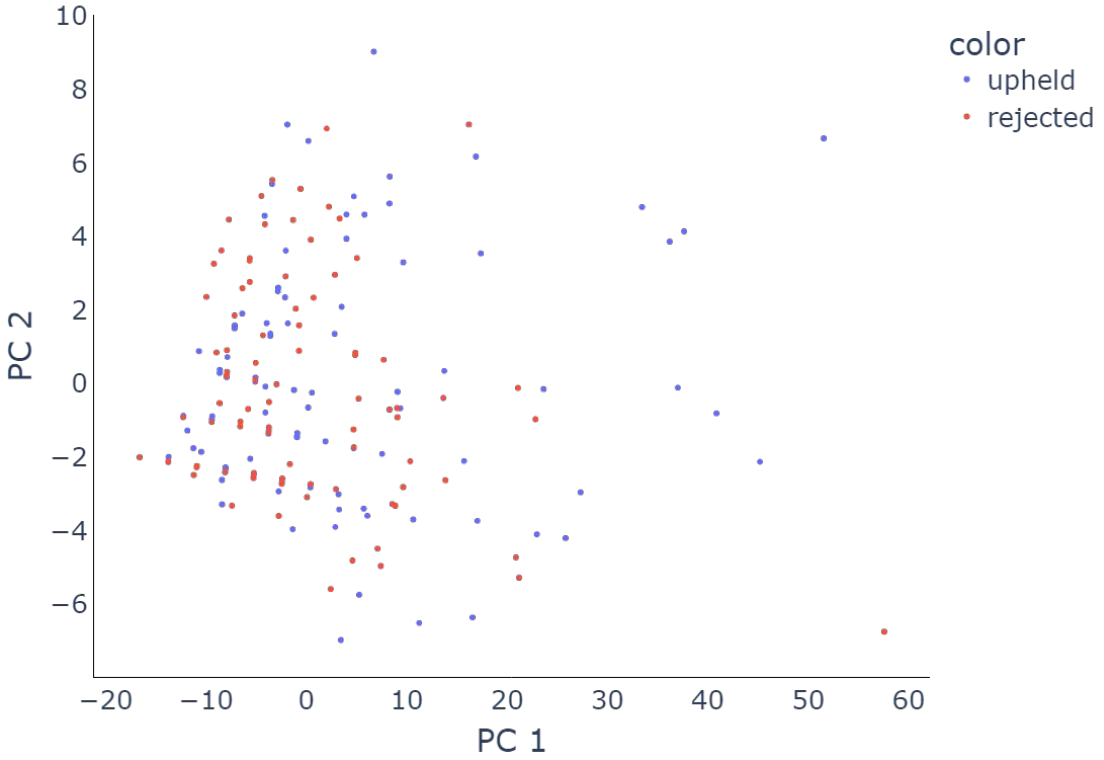


Figure 6.3: Scatterplot of the first two Principal Components of Features concerning Visual Similarity of Word Marks

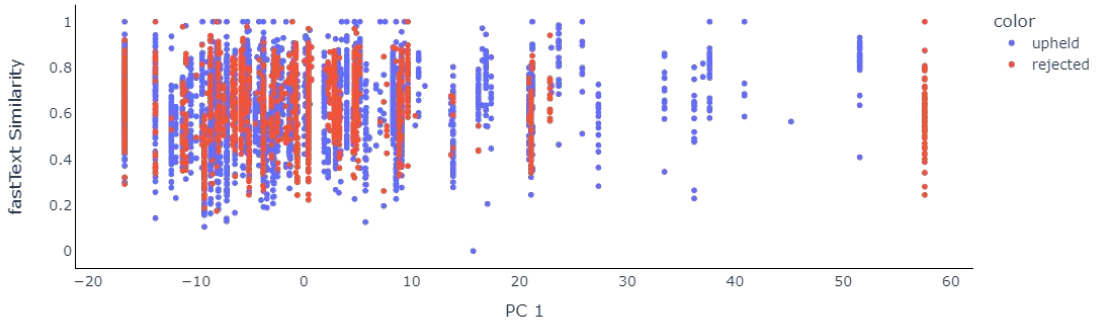


Figure 6.4: Scatterplot of the first Principal Component of Features concerning Visual Similarity of Word Marks and the respective Similarity of Goods and Services computed with fastText

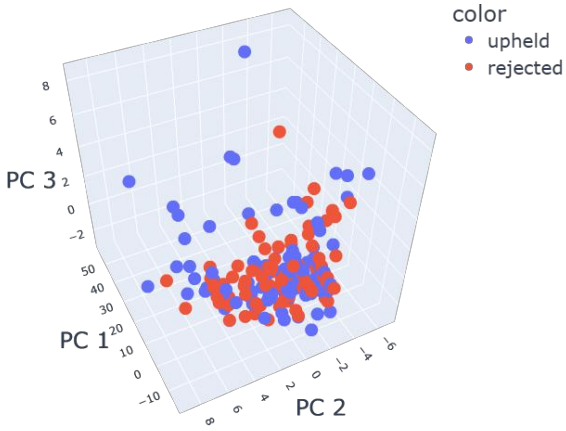


Figure 6.5: Scatterplot of the first three Principal Components of Features concerning the Visual Similarity of Word Marks

Aural similarity is computed using a combination of a phonetic encoding and a string similarity measure. Since phonetic encodings strictly follow substitution rules, the correlation between features for aural similarity is also high. The correlations for these features are shown in figure 6.6. The correlation matrix does not show a significant difference between the two phonetic encodings. Even aural similarities of different phonetic encodings correlate with each other. The reason for this observation is that the double metaphone algorithm and the metaphone 3 algorithm have similar substitution rules which result in a similar phonetic encoding for most trademarks.

Models

TrademarkML uses RFs and SVMs with a linear kernel as these models are performant while also being interpretable [142, p 4831] [115, p 1429]. SVMs were restricted to a linear kernel for performance reasons. As the number of training samples increases, repeated hyperparameter tuning for non-linear kernels becomes unfeasible. Both classifiers are used with a seed of 42. Furthermore, the linear SVM is used with a parameter that allows the model to automatically choose whether to solve the dual or the primal optimization problem. Other parameters are subject to the hyperparameter tuning process.

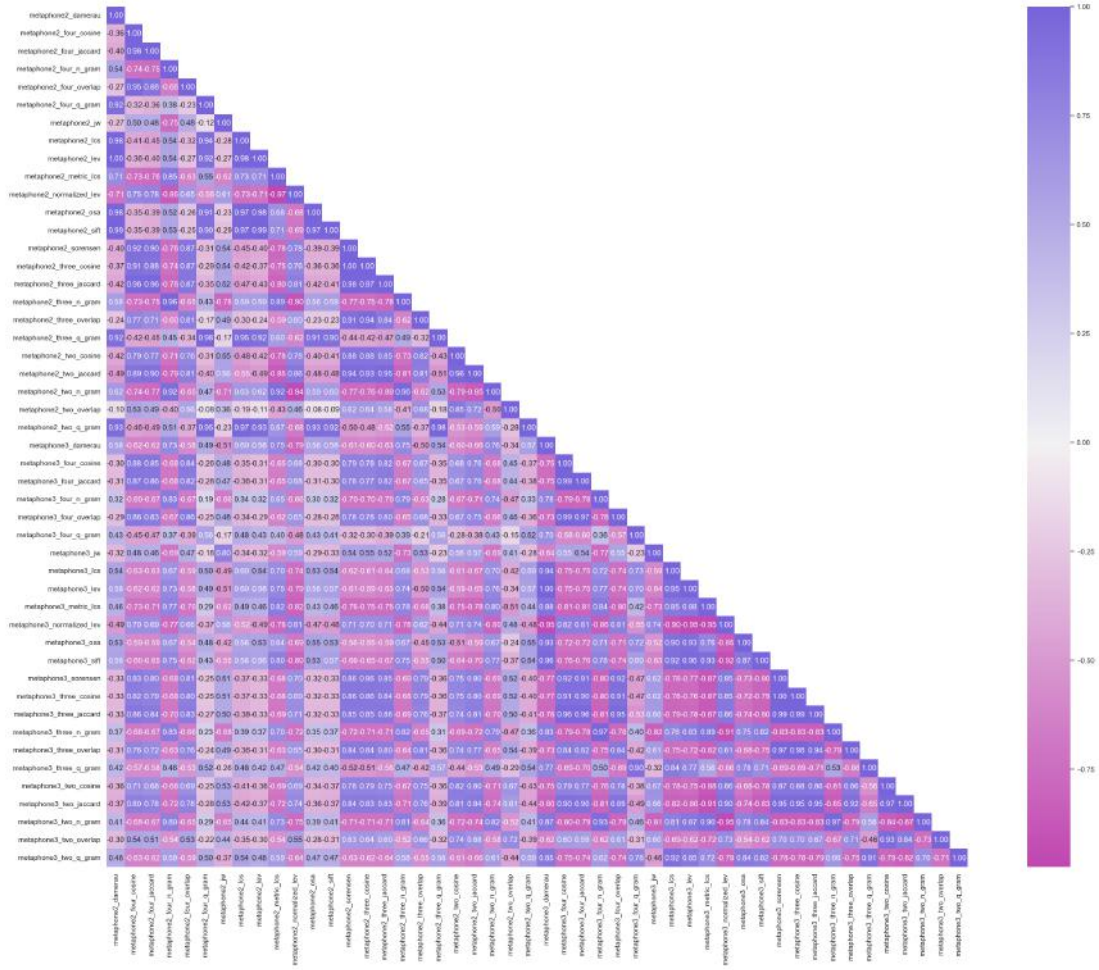


Figure 6.6: Correlation Matrix for Features concerning Aural Similarity

Data Preprocessing

For all experiments, the target variable is binarized so that it can be used for binary classification. The labels are transformed into values 0 and 1, depending on their string value.

Since the RF classifier is scale-invariant, there is no need to scale the features beforehand. For experiments using an SVM, however, scaling is performed. Each experiment is carried out three times to compare different scaling methods. This increases the number of models trained for SVM by a factor of three as can be seen in table 6.3.

Feature Selection

In order to find the best feature combination to predict the likelihood of confusion, an exhaustive comparison is performed between feature groups. This comparison also allows to examine the influence of single features on the prediction performance. For each combination of features, a model is trained from scratch and tuned to the respective subset of the data. Then, its performance is evaluated on the independent test set. As the features measuring visual and aural similarities were found to be highly correlated within each group, the first principal component of each group was also used as a possible feature in the exhaustive feature selection process.

	Random Forest		**Support Vector Machine**	
	Word	**Figurative**	**Word**	**Figurative**
#Models	15,911	3,059	47,733	9,177

Table 6.3: Number of Models trained

Hyperparameter Tuning

Hyperparameter tuning is performed for each combination of features using the GridSearchCV function by sklearn in combination with GroupShuffleSplit so that two splits do not contain samples of the same opposition case. A 5-fold CV is performed. Estimators are evaluated and compared on the validation set created during CV based on their prediction accuracy. Each model has their own parameter grid.

For the RF classifier, the grid contains values for the maximum depth of each decision tree (25, 50, or 75), the maximum number of features (log2 or sqrt), and the number of decision trees (15, 20, or 50). The number of parameters for the linear SVM is much lower. For the SVM only the regularization parameter C is tuned using the values 0.01, 0.1, 1, 10, and 100.

CHAPTER 7

Results

In this chapter, results from the extensive evaluation performed from the prediction module of TrademarkML are analyzed in two ways. The quantitative analysis reports the numbers and compares methods solely based on their measured performance. The qualitative analysis, on the other hand, focusses on aspects of the performance that the quantitative analysis does not capture.

7.1 Quantitative Analysis

The experiments return four files, one file per classifier and type of trademark. The first line of each file contains a reference to the best iteration based on the model's prediction accuracy on the test set. The performances reached by the best iterations are listed in table 7.1.

Performance Metric	**Model**			
	Word Mark Data		**Figurative Mark Data**	
	RF	**SVM**	**RF**	**SVM**
F1 Score	0.88	0.88	0.81	0.74
Accuracy	0.82	0.82	0.78	0.71
Precision	0.80	0.80	0.77	0.73
Recall	0.97	0.98	0.84	0.75
ROC AUC	0.73	0.73	0.77	0.71

Table 7.1: Performance of Classifiers with the highest Accuracy on the Test Set

For word marks, there is no significant difference between the two classifiers, RF and SVM. For figurative marks, however, the RF classifier outperforms the SVM by 6.8% on

average. Each iteration also returns the subset of features used. The features found for the best performing models are shown in table 7.2.

Model	**Features**
RF (word)	Q-gram(n=2), Metaphone 3 + Cosine(n=4)
SVM (word)	Sorensen, Metaphone 3 + LCS, WordNet + Cosine
RF (figurative)	ResNet50, Double Metaphone + Cosine(n=2), WordNet + Cosine
SVM (figurative)	VGG19, Metaphone 3 + Cosine(n=2), WordNet + Levenshtein, Google Universal Sentence Encoder, Normalized

Table 7.2: Features leading to the best Accuracy per Model

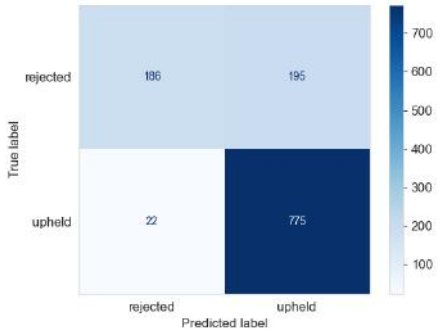


Figure 7.1: Confusion Matrix: RF for Word Marks

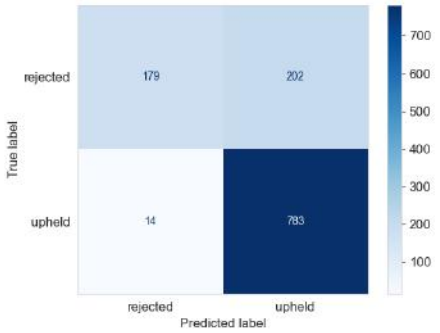


Figure 7.2: Confusion Matrix: SVM for Word Marks

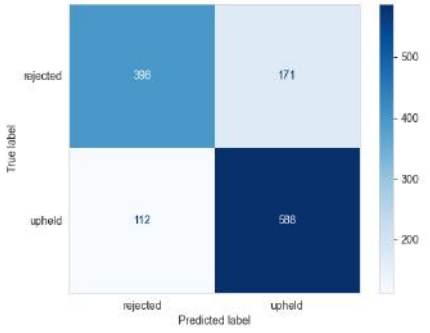


Figure 7.3: Confusion Matrix: RF for Figurative Marks

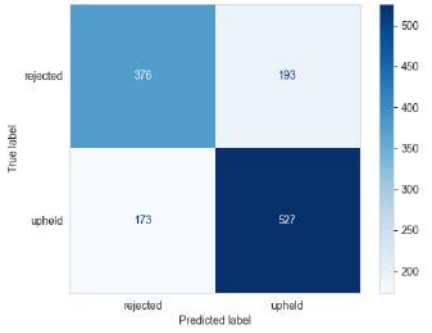


Figure 7.4: Confusion Matrix: SVM for Figurative Marks

Confusion matrices can be constructed using the ground truth and the predicted labels. Figures 7.1 and 7.2 show the confusion matrices of the two tuned models on word mark

data. Figures 7.3 and 7.4 show the same metric for figurative mark data. ROC-Curves and precision-recall-curves are shown in appendix 9.

7.2 Qualitative Analysis

As can be seen in table 7.2, the chosen models were more performant on word mark data. However, no significant difference in performance is observed between the RF and the SVM classifier on word mark data. For figurative marks, on the other side, the performance strongly depends on the model. The RF classifier outperforms the SVM in all aspects by 6.8% on average.

For both use cases, trademark monitoring and similarity search for earlier trademarks, the recall of the final model is the most important metric, as false negatives can lead to legal consequences. false positives, on the other hand, lead to higher costs, as manual work and expert knowledge is required to distinguish relevant from irrelevant results. Furthermore, relying on false positives can lead to procedural costs.

7.2.1 Models for Word Mark Data

Figures 7.1 and 7.2 show that the SVM is more likely to predict the class label "upheld" than the RF on word mark data. Since recall is the most important performance metric for the previously defined use cases, the SVM is considered to perform slightly better than the RF. However, both models are biased towards positive predictions, leading to a high number of false positives. The recall-precision curves are shown in the figures A1 and A2. With an average precision of 80%, the RF yields a slightly better ratio of true positives to false positives.

The learning curves of both models are visualized in figures A5 and A6. Both plots show that the training accuracy is much higher than the validation accuracy. This means that both models are overfitting. While the validation accuracy slightly increases with sample size in figure A5, the validation accuracy stagnates at a validation set size of around 60%. After hyperparameter optimization, both models, however, show a higher test accuracy than validation accuracy.

RFs allow to inspect the feature importance. For the best performing RF on word mark data, the feature importance is shown in figure 7.5. Interestingly, the feature for visual similarity has an importance of over 84% while the feature for aural similarity contributes to not even 16%.

Furthermore, the linear SVM also allows to inspect the feature importance by analyzing its weights [61, p 7 ff]. The magnitude of a feature's weight corresponds to its importance. For word mark data, the optimized SVM has the weights

$$\boldsymbol{w} = [0.1369,\ -0.0327,\ 0.4500], \tag{7.1}$$

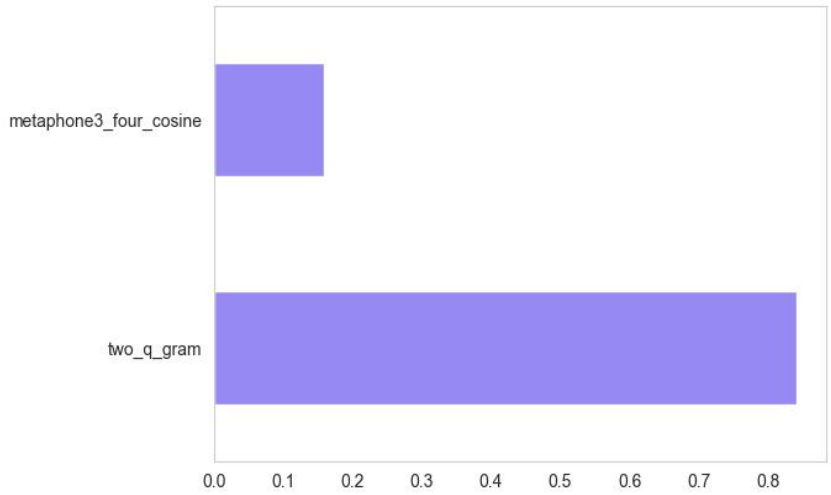


Figure 7.5: RF for Word Mark Data: Feature Importance

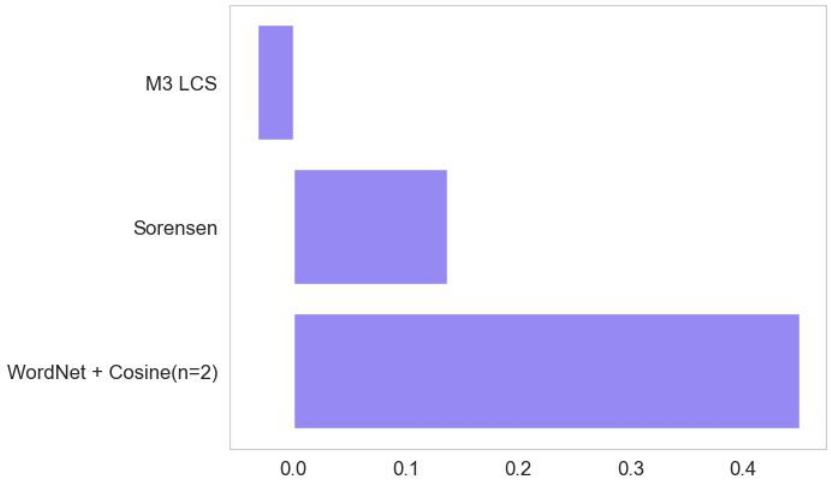


Figure 7.6: SVM for Word Mark Data: Coefficients

which are visualized in figure 7.6.

Interestingly, the conceptual similarity of trademarks is the most important feature for the SVM classifier. This observation is not in line with the ground truth, as conceptual similarity often can not even be assessed and, therefore, does not contribute to the global assessment. However, conceptual similarity might be correlated to visual similarity in some cases, like “SHAMAN” and “SHAMAN’S” in opposition case 003179493.

On the test set, the prediction of the two models differ in six opposition cases with a

total of 171 contested goods and services. Interestingly, for all those cases, the aural similarity used for the RF classifier is 0.0.

Both models, RF and SVM, do not consider the similarity of goods and services. This makes it impossible to make correct predictions for partially upheld oppositions. For example, both classifiers fail to correctly predict opposition case 003159285. This case deals with seven goods and services that conflict with the earlier marks' goods and services and nine goods and services for which no likelihood of confusion exists.

Also, the RF classifier shows unexpected behaviour for some instances. Opposition case 003097006 deals with the trademarks "JOY" and "PROFUMI DI PANTELLERIA JOYANN". For this case, the visual distance is 26 and the aural similarity is 0.0, using the q-gram feature with n=2 and the metaphone 3 phonetic encoding with a cosine similarity with n=4. Even though these features indicate a very high distance, the model predicts that a likelihood of confusion is given. While it is clear that these two features alone cannot lead to satisfactory results, a higher distance between trademarks should not lead to the prediction that there is a likelihood of confusion if a smaller distance can yield the opposite result.

Figures A9, A10, and A11 show the decision boundaries for the optimized SVM model for word mark data. These boundaries indicate that the model is biased towards positive predictions. The boundary between visual and aural similarity, shown in figure A9, could be improved by making it more sensitive to changes in the visual similarity. The other two decision boundaries seem to work well for test data, considering that the data cannot be linearly separated.

7.2.2 Models for Figurative Mark Data

As shown in table 7.1, the model selection makes a significant difference for predicting figurative mark data. The SVM classifier predicts both more false positives and more false negatives on the test set.

Intestingly, both models consider more features than when optimized for word mark data. Figures 7.7 and 7.8 show the feature importance for each model. The RF is the only model that uses the double metaphone algorithm.

The SVM is the only model that utilizes a feature that measures the similarity of goods and services. Interestingly, its feature for visual similarity has the lowest impact on its predictions and its feature for aural similarity has the strongest impact. The coefficients are

$$\boldsymbol{w} = [0.0661, 0.5889, 0.4431, 0.1897]. \tag{7.2}$$

Figures A7 and A8 show the learning curves for both models. The training accuracy for the RF classifier is much higher than its validation accuracy for any sample size. The model is therefore heavily overfitting. The SVM also has a higher training accuracy.

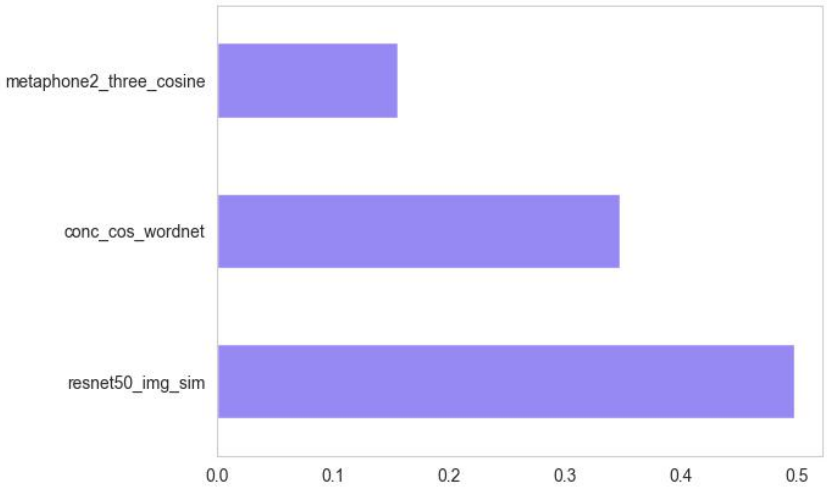


Figure 7.7: RF for Figurative Mark Data: Feature Importance

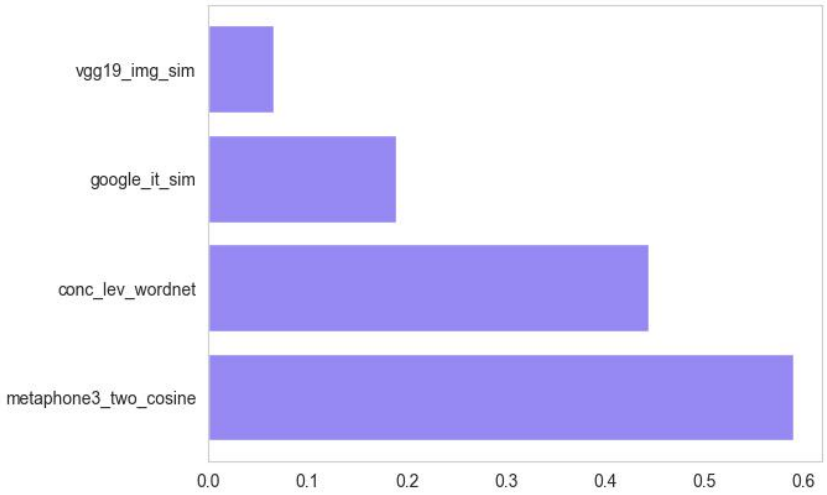


Figure 7.8: SVM for Figurative Mark Data: Coefficients

However, the gap between training accuracy and validation accuracy is not as big as for the RF.

Figures A12, A13, A14, A15, A16, and A17 show the decision boundaries for the optimized SVM model for figurative mark data. In contrast to the decision boundaries in appendix 9, this SVM model considers the similarity of goods and services. This allows to inspect the decision boundary along that variable.

In fact, the SVM classifier for figurative mark data is the only model that can predict likelihood of confusion for each single contested good or service. For case number 003170664, the model predicted a likelihood of confusion for all contested goods and services except for "preparations based on cereals", as its similarity to the earlier trademark's "Flour and preparations based on cereals; bread, pastry and confectionery products; yeast, baking powders" is 81% using the Google's universal sentence encoder. However, for "preparations based on cereals" the opposition is actually upheld. For other samples from case number 003170664, namely "aromatic preparations for pastries" and "flavourings for cakes", the model predicts the existence of likelihood of confusion, even though the opposition is rejected for these goods and services.

7.2.3 Impact of Feature Selection

The exhaustive feature selection process in these experiments allow to compare the performance of each feature separately within each feature group. The average F1-score of each feature and model is shown in appendix 9.

However, there is no pattern in the data. Instead, there are many interesting findings that do not seem to be correlated between experiments.

Word Mark Data

Most features for visual similarity yield an average F1-score of 70% to 80% for both models, RFs and SVMs. The Szymkiewicz-Simpson coefficient with n=2 and n=3, however, lead to a lower average F1-score of 65% to 68% for SVMs. For both models, the Q-gram similarity achieves the best performance.

For RFs, no significant differences in the average performance between aural similarity features can be observed. Interestingly, the first principal component of all features using the metaphone 3 algorithm achieves, on average, the second best F1-score and outperforms each single feature computed using the metaphone 3 algorithm. However, the best performing feature is the Jaccard index with n=2 on the phentic encoding produced by the double metaphone algorithm.

SVMs, on the other hand, are slightly more sensitive to the selection of aural similarity features. Most features achieve an average F1-score of 80%. Analogous to the visual similarity features, the combination of the Szymkiewicz-Simpson coefficient and the metaphone 3 algorithm yields the lowest average performance. The best average performance is achieved using the Q-gram similarity with n=4 and the metaphone 3 algorithm.

For both models, the selection of conceptual similarity feature does not have a strong impact on the prediction performance. For SVMs, the average performance is actually best without any features of that feature group. RFs, on the other hand, perform best using cosine similarity to map trademark names to the WordNet ontology.

For RFs, there is no significant difference between Google's universal sentence encoder and fastText. Also, these two features do not decrease the variance of the average prediction

performance. However, without these features, many more outliers on the lower end can be observed. This is different for SVMs. Google's universal sentence encoder decreases the average performance significantly, while fastText and omitting features of this feature group achieve the same average F1-score of almost 80%.

Figurative Mark Data

Interestingly, the feature extracted from the ResNet50 leads to the best performance in RFs but performs worst for SVMs. On average, SVMs perform best without any image similarity feature. However, both, the VGG16 and the VGG19 model, return a visual similarity that yields a performance close to the performance achieved without any visual similarity features. In fact, in some iterations, the VGG19 feature performs best. For this reason, it is used for the optimized model, as shown in table 7.1. This means that, generally, image similarity features introduce non-linearity, which cannot be handled by an SVM with a linear kernel. In combination with other features, however, linear separability can be improved. For both models, the first principal component of all features related to visual similarity yields a relatively poor performance, compared to other features.

Another interesting observation is that for RFs there is no significant difference in the features for aural similarity. Still, on average, features using the double metaphone algorithm achieve the best performance. For SVMs, however, the metaphone 3 algorithm leads to a significantly better F1-score.

For both, SVMs and RFs, the cosine similarity and the LCS perform best for mapping trademarks to the WordNet ontology. These features lead to a significant improvement for SVMs. However, for RFs, omitting these features does not have a strong impact on the average F1-score.

Google's universal sentence encoder and fastText improve the average performance of RFs and SVMs. For both models, fastText yields an average performance with a much lower variance than using Google's universal sentence encoder or omitting features of this feature group.

7.2.4 Bias and Limitations

The selection process of the data imposes selection bias in the data. This means that the models are trained only on trademarks that were similar enough to be subject to opposition cases. However, the use case of the model is also to compare completely different trademarks. Furthermore, the data only consists of 500 opposition cases. This sample size is high compared to the sample size used in related works. However, compared to the variance expected in production data, the sample size is relatively low.

The resulting models are limited to the features listed in table 6.2. Different models could improve the models' performances. The same applies for scaling methods employed for SVMs. Also, regardless of the method used for mapping trademarks to the ontology,

the features for conceptual similarity strongly depend on the WordNet ontology. Thus, using another ontology could also improve the models' predictions.

Another limitation is that the models are only applicable to marks with a textual representation. Figurative marks without a name cannot be processed, as a character sequence is needed for the phonetic comparison.

Overall, the optimized models are biased towards false positives. Consequently. the vast majority of trademarks creating a likelihood of confusion is detected. However, around 20% of these predictions are wrong. Still, these models can help reduce the number of comparisons needed for identifying conflicting trademarks in the task of trademark monitoring.

CHAPTER 8

Conclusion

Many different methods to compare trademarks on a visual, aural, and conceptual level were developed over time. However, existing solutions to compare trademarks do not take into account the similarity of goods and services. The current literature does not assess datasets that contain partially upheld positions. If such datasets were taken into account, they would show the necessity of considering this similarity aspect.

Using state-of-the-art methods from existing literature and a novel approach for computing the similarity of goods and services using text embeddings, the prediction of likelihood of confusion works reasonably well on trademark level. The best models found in this thesis achieve an F1-score of 88% for word marks and 81% for figurative marks on independent test data. The even higher recall of 98% for word marks and 84% for figurative marks means that false negatives are rather unlikely. Furthermore, a precision of 80% for word marks and 77% for figurative marks is achieved.

For word mark data, the optimized SVM achieves the best performance by making predictions based on the Sorensen-Dice coefficient, the LCS on the phonetic encoding of the trademark using the metaphone 3 algorithm, and the Wu-Palmer similarity between nodes in the WordNet ontology, to which the trademarks were mapped using cosine similarity.

For figurative marks, on the other hand, RF outperforms SVM. The best performance is achieved by using the similarity extracted with the ResNet50 model, the cosine similarity on the phonetic encoding of the trademark using the double metaphone algorithm, the Wu-Palmer similarity between nodes in the WordNet ontology, to which the trademarks were mapped using Levenshtein distance, and the similarity of goods and services, computed using Google's universal sentence encoder.

Overall, the classifiers perform better on word mark data by 6.8% on average for each metric. Interestingly, top-performing models do not make use of the similarity of goods and services, even though partially upheld oppositions are contained in the dataset.

The results show the need for a larger dataset and more meaningful features. Even though the performance achieved by TrademarkML is quite good, the qualitative analysis proves that the basis on which the predictions on trademark level are made are not completely in line with the examination guidelines for likelihood of confusion, as the predictions for word marks mainly depend on the conceptual similarity of the marks.

Future research is therefore needed to establish a data mining process for trademark opposition decisions and develop new methods for extracting meaningful features.

CHAPTER 9

Summary

Over the past years, the number of trademark application has risen continuously. To make it easier for trademark owners to protect their trademark's territory and for trademark applicants to know if their trademark is likely to be refused, an automated and reliable tool is needed to compare two trademarks and their respective goods and services. Most of the existing services offering a similarity search just consider the spelling of the trademarks names, which is insufficient to assess likelihood of confusion.

For this reason, this thesis introduces the concept of a trademark management system, TrademarkML. TrademarkML faciliates tasks like trademark monitoring and searching for conflicting trademarks by automatically classifying trademark pairs.

This thesis attempts to implement and evaluate the prediction module of TrademarkML. To train the models used by TrademarkML, the dataset TMSIM-500 is used. This dataset consists of data from 500 opposition cases taken from the EUIPO database [39]. The predictions are based on state-of-the-art features known from existing literature. For measuring the visual similarity of two trademarks, string comparison methods, such as the Levenshtein distance or cosine similarity, are employed for word marks and features extracted by CNNs are used for figurative marks. Aural similarity is computed by using the same string similarity methods as for computing visual similarity on the phonetic encoding of the trademarks' names. Conceptual similarity is derived by mapping the trademarks' names to nodes of the WordNet ontology using three string comparison methods - LCS, cosine, and Levenshtein - and then computing the Wu-Palmer similarity between the two nodes. The similarity between goods and services is computed using Google's universal sentence encoder and fastText.

RFs and SVMs with a linear kernel are then optimized by performing an exhaustive feature selection that iterates over all feature combinations. For each feature combination, the model's hyperparameters are tuned on a 5-fold CV using 80% of the data. Each fold uses 80% of that data for training data and the remaining data as validation set.

The results show that for word mark data the optimized SVM achieves the best performance by making predictions based on the Sorensen-Dice coefficient, the LCS on the phonetic encoding of the trademark using the metaphone 3 algorithm, and the Wu-Palmer similarity between nodes in the WordNet ontology, to which the trademarks were mapped using cosine similarity. For figurative marks, on the other hand, RF outperforms SVM. The best performance is achieved by using the similarity extracted with the ResNet50 model, the cosine similarity on the phonetic encoding of the trademark using the double metaphone algorithm, the Wu-Palmer similarity between nodes in the WordNet ontology, to which the trademarks were mapped using Levenshtein distance, and the similarity of goods and services, computed using Google's universal sentence encoder.

The best models achieve an F1-score of 88% for word marks and 81% for figurative marks, a recall of 98% for word marks and 84% for figurative marks, and a precision of 80% for word marks and 77% for figurative marks. Overall, classifiers perform better on word mark data by 6.8% on average for each metric.

None of the top-performing models uses a feature for meaning the similarity of goods and services. This makes it impossible to correctly predict partially upheld oppositions. A larger dataset and more meaningful features are required to overcome this issue.

List of Figures

List of Tables

List of Algorithms

Glossary

European Union Trade Marks Bulletin A publication by the EUIPO containing the latest trademark applications. 2, 55

Acronyms

CBIR content-based image retrieval. 51

CBOW Continuous-Bag-of-Words. 28, 29

CJEU Court of Justice of the European Union. 3, 7, 9, 11, 12, 17

CNN convolutional neural network. 31, 54, 57, 77

CV cross-validation. 25, 26, 64, 77

EU European Union. 5, 8, 10

EUIPO European Union Intellectual Property Office. 2, 3, 6, 12–17, 37, 77

EUTM European Union Trademark. 6, 8

GIs Geographical Indications. 10

GloVe Global Vectors. 29

LCS longest common substring. 29, 30, 53, 56–58, 72, 75, 77, 78

OCR Optical Character Recognition. 30

OHIM Office for Harmonization in the Internal Market. 2, 6

PCA Principal Component Analysis. 59

PVD Plant Variety Denomination. 10

RF random forest. 20, 23, 24, 62–72, 75, 77–80, 117–120, 126–129

SVM support vector machine. 20, 22, 23, 62–72, 75, 77–81, 117–126, 130–134

TF-IDF term frequency-inverse document frequency. 31

TTWs Traditional Terms of Wine. 10

Legal Texts

EUTMDR Commission Delegated Regulation (EU) 2018/625 of 5 March 2018 supplementing Regulation (EU) 2017/1001 of the European Parliament and of the Council on the European Union trade mark, and repealing Delegated Regulation (EU) 2017/1430, OJ L 2018/104

EUTMIR Commission Implementing Regulation (EU) 2018/626 of 5 March 2018 laying down detailed rules for implementing certain provisions of Regulation (EU) 2017/1001 of the European Parliament and of the Council on the European Union trade mark, and repealing Implementing Regulation (EU) 2017/1431, OJ L 2018/104

EUTMR Regulation (EU) 2017/1001 of the European Parliament and of the Council of 14 June 2017 on the European Union trade mark, OJ L 2017/154

Lisbon Treaty Treaty of Lisbon amending the Treaty on European Union and the Treaty establishing the European Community, signed at Lisbon, 13 December 2007, OJ C 2007/306

Nice Classification World Intellectual Property Organization, International Classification of Goods and Services for the Purposes of the Registration of Marks (Nice Classification), Eleventh Edition, 2022

Paris Convention Paris Convention for the Protection of Industrial Property of March 20, 1883, as revised at Brussels on December 14, 1900, at Washington on June 2, 1911, at The Hague on November 6, 1925, at London on June 2, 1934, at Lisbon on October 31, 1958, and at Stockholm on July 14, 1967, and as amended on September 28, 1979

Regulation (EU) 1151/2012 Regulation (EU) 1151/2012 of the European Parliament and of the Council of 21 November 2012 on quality schemes for agricultural products and foodstuffs, OJ L 2012/343

Regulation (EU) 1308/2013 Regulation (EU) 1308/2013 of the European Parliament and of the Council of 17 December 2013 establishing a common organisation of the markets in agricultural products and repealing Council Regulations (EEC) No 922/72, (EEC) No 234/79, (EC) No 1037/2001 and (EC) No 1234/2007, OJ L 2013/347

Regulation (EU) 2015/2424 Regulation (EU) 2015/2424 of the European Parliament and of the Council of 16 December 2015 amending Council Regulation (EC) No 207/2009 on the Community trade mark and Commission Regulation (EC) No 2868/95 implementing Council Regulation (EC) No 40/94 on the Community trade mark, and repealing Commission Regulation (EC) No 2869/95 on the fees payable to the Office for Harmonization in the Internal Market (Trade Marks and Designs), OJ L 2015/341

Regulation (EU) 2019/787 Regulation (EU) 2019/787 of the European Parliament and of the Council of 17 April 2019 on the definition, description, presentation and labelling of spirit drinks, the use of the names of spirit drinks in the presentation and labelling of other foodstuffs, the protection of geographical indications for spirit drinks, the use of ethyl alcohol and distillates of agricultural origin in alcoholic beverages, and repealing Regulation (EC) No 110/2008, OJ L 2019/130

Vienna Classification World Intellectual Property Organization, International Classification of the Figurative Elements of Marks under the Vienna Agreement (Vienna Classification), Ninth Edition, 2022, in force as from January, 2023

Bibliography

[1] Apurva Agarwal, Deepti Agrawal, and Dilip Kumar Sharma. Trademark image retrieval using color and shape features and similarity measurement. In *2021 8th International Conference on Computing for Sustainable Global Development (INDIACom)*, pages 486–490, 2021.

[2] Deepti Agrawal, Anand Singh Jalal, and Rajesh Tripathi. Trademark image retrieval by integrating shape with texture feature. In *2013 International Conference on Information Systems and Computer Networks*, pages 30–33, 2013. doi: 10.1109/ICISCON.2013.6524168.

[3] Akiko Aizawa. The feature quantity: An information theoretic perspective of tfidf-like measures. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '00, page 104111, New York, NY, USA, 2000. Association for Computing Machinery. ISBN 1581132263. doi: 10.1145/345508.345556. URL https://doi.org/10.1145/345508.345556.

[4] Hayfa Alshowaish, Yousef Al-Ohali, and Abeer Al-Nafjan. Trademark image similarity detection using convolutional neural network. *Applied Sciences*, 12(3), 2022. ISSN 2076-3417. doi: 10.3390/app12031752. URL https://www.mdpi.com/2076-3417/12/3/1752.

[5] Alwis and Austin. An integrated framework for trademark image retrieval using gestalt features and cmm neural network. In *Image Processing And Its Applications, 1999. Seventh International Conference on (Conf. Publ. No. 465)*, volume 1, pages 290–295 vol.1, 1999. doi: 10.1049/cp:19990329.

[6] S. Alwis and J. Austin. Searching image databases containing trademarks. In *IEE Colloquium on Neural Networks in Interactive Multimedia Systems (Ref. No. 1998/446)*, pages 2/1–2/5, 1998. doi: 10.1049/ic:19980710.

[7] Gidudu Anthony, Hulley Gregg, and Marwala Tshilidzi. Image classification using svms: One-against-one vs one-against-all, 2007.

[8] Fatahiyah Mohd Anuar, Rossitza Setchi, and Yu-Kun Lai. A conceptual model of trademark retrieval based on conceptual similarity. *Procedia Computer Science*, 22: 450–459, 2013.

[9] Fatahiyah Mohd Anuar, Rossitza Setchi, and Yu-Kun Lai. Semantic retrieval of trademarks based on conceptual similarity. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 46(2):220–233, 2016. doi: 10.1109/TSMC.2015.2421878.

[10] Sylvain Arlot and Alain Celisse. A survey of cross-validation procedures for model selection. *Statistics Surveys*, 4(none), jan 2010. doi: 10.1214/09-ss054. URL `https://doi.org/10.1214%2F09-ss054`.

[11] Mariette Awad and Rahul Khanna. *Support Vector Machines for Classification*, pages 39–66. 04 2015. ISBN 978-1-4302-5989-3. doi: 10.1007/978-1-4302-5990-9_3.

[12] Daniel Bakkelund. An lcs-based string metric. 2009. URL `https://api.semanticscholar.org/CorpusID:5116711`.

[13] Rémi Bardenet, Mátyás Brendel, Balázs Kégl, and Michèle Sebag. Collaborative hyperparameter tuning. In Sanjoy Dasgupta and David McAllester, editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 199–207, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR. URL `https://proceedings.mlr.press/v28/bardenet13.html`.

[14] Ann Bartow. Likelihood of confusion. *San Diego L. Rev.*, 41:721, 2004.

[15] Ann Bartow. Exporting trademark confusion. In *Intellectual Property Rights in a Networked World: Theory and Practice*, pages 113–160. IGI Global, 2005.

[16] Barton Beebe. An empirical study of the multifactor tests for trademark infringement. *Calif. L. Rev.*, 94:1581, 2006.

[17] Kevin Blum. Consistency or confusion: A fifteen-year revisiting of barton beebe's empirical analysis of multifactor tests for trademark infringement. *Stan. Tech. L. Rev.*, page 3, 2010.

[18] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*, 2016.

[19] Robert G Bone. Taking the confusion out of likelihood of confusion: Toward a more sensible approach to trademark infringement. *Nw. UL Rev.*, 106:1307, 2012.

[20] Mary Fran Buehler. Report construction: Tables. *IEEE Transactions on Professional Communication*, PC-20(1):29–32, 1977. doi: 10.1109/TPC.1977.6594173.

[21] Mateo Cano-Solis, John R. Ballesteros, and John W. Branch-Bedoya. Vepl dataset: A vegetation encroachment in power line corridors dataset for semantic segmentation of drone aerial orthomosaics. *Data*, 8(8), 2023. ISSN 2306-5729. doi: 10.3390/data 8080128. URL https://www.mdpi.com/2306-5729/8/8/128.

[22] Daniel Cer, Yinfei Yang, Sheng yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. Universal sentence encoder, 2018.

[23] Abdolah Chalechale and Amir Faramarzi. Incorporating efficiency and human judgment in image retrieval for trademark matching. In *2010 6th Iranian Conference on Machine Vision and Image Processing*, pages 1–6, 2010. doi: 10.1109/IranianM VIP.2010.5941135.

[24] J. Charles. Cmu's autonomous helicopter explores new territory. *IEEE Intelligent Systems and their Applications*, 13(5):85–87, 1998. doi: 10.1109/5254.722385.

[25] Cai-kou Chen, Qiang-qiang Sun, and Jing-yu Yang. Binary trademark image retrieval using region orientation information entropy. In *2007 International Conference on Computational Intelligence and Security Workshops (CISW 2007)*, pages 295–298, 2007. doi: 10.1109/CISW.2007.4425495.

[26] Hong Chen. Trademark detection algorithm based on artificial intelligence. In *2023 4th International Conference for Emerging Technology (INCET)*, pages 1–6, 2023. doi: 10.1109/INCET57972.2023.10169971.

[27] Seung-Seok Choi, Sung-Hyuk Cha, Charles C Tappert, et al. A survey of binary similarity and distance measures. *Journal of systemics, cybernetics and informatics*, 8(1):43–48, 2010.

[28] G. Ciocca and R. Schettini. Similarity retrieval of trademark images. In *Proceedings 10th International Conference on Image Analysis and Processing*, pages 915–920, 1999. doi: 10.1109/ICIAP.1999.797712.

[29] Ben Coffey. General court confirms that the body shop cannot register wisdom spaas an eu trade mark. *Journal of Intellectual Property Law & Practice*, 11(8): 570–572, 2016.

[30] Cahyo Crysdian. Content based image retrieval system based on watershed transform for trademark images. In *2014 Electrical Power, Electronics, Communicatons, Control and Informatics Seminar (EECCIS)*, pages 116–120, 2014. doi: 10.1109/EECCIS.2014.7003730.

[31] Fred J. Damerau. A technique for computer detection and correction of spelling errors. *Commun. ACM*, 7(3):171176, mar 1964. ISSN 0001-0782. doi: 10.1145/36 3958.363994. URL https://doi.org/10.1145/363958.363994.

[32] John P Eakins, Jago M Boardman, and Margaret E Graham. Similarity retrieval of trademark images. *IEEE multimedia*, 5(2):53–63, 1998.

[33] John P Eakins, K Jonathan Riley, and Jonathan D Edwards. Shape feature matching for trademark image retrieval. In *Image and Video Retrieval: Second International Conference, CIVR 2003 Urbana-Champaign, IL, USA, July 24–25, 2003 Proceedings 2*, pages 28–38. Springer, 2003.

[34] Chuck Easttom. On the use of the ssim algorithm for detecting intellectual property copying in web design. In *2021 IEEE 11th Annual Computing and Communication Workshop and Conference (CCWC)*, pages 0860–0864, 2021. doi: 10.1109/CCWC51732.2021.9375928.

[35] Mostafa El Habib Daho, Nesma Settouti, Mohammed El Amine Lazouni, and Mohammed El Amine Chikh. Weighted vote for trees aggregation in random forest. In *2014 International Conference on Multimedia Computing and Systems (ICMCS)*, pages 438–443, 2014. doi: 10.1109/ICMCS.2014.6911187.

[36] Safet Emruli, Agim Nuhiu, and Besa Kadriu. Trademark protection, absolute and relative grounds for refusal of trademark. *European Journal of Multidisciplinary Studies*, 1(2):291–297, 2016.

[37] Tobias Endrich-Laimböck and Svenja Schenk. Then tell me what you think about morality: A freedom of expression perspective on the cjeus decision in fack ju goĺhte (c-240/18 p). *International Review of Intellectual Property and Competition Law*, 51: 529–542, 2020. URL `https://doi.org/10.1007/s40319-020-00936-9`.

[38] B. Erol and F. Kossentini. A robust distance measure for the retrieval of video objects. In *Proceedings Fifth IEEE Southwest Symposium on Image Analysis and Interpretation*, pages 40–44, 2002. doi: 10.1109/IAI.2002.999886.

[39] EUIPO. eSearch Case Law. `https://euipo.europa.eu/eSearchCLW/`, 2023. Accessed: 23 August 2023.

[40] EUIPO. eSearch plus. `https://euipo.europa.eu/eSearch`, 2023. Accessed: 24 August 2023.

[41] EUIPO. Basic questions. `https://euipo.europa.eu/ohimportal/en/eutm-general-questions`, 2023. Accessed: 24 August 2023.

[42] EUIPO. Statistics for european union trade marks: 1996-01 to 2023-07 evolution. `https://euipo.europa.eu/tunnel-web/secure/webdav/guest/document_library/contentPdfs/about_euipo/the_office/statistics-of-european-union-trade-marks_en.pdf`, 2023. Accessed: 24 August 2023.

[43] EUIPO. Trade marks in the european union. https://euipo.europa.eu/ohimportal/en/trade-marks-in-the-european-union, 2023. Accessed: 25 August 2023.

[44] EUIPO. Guidelines for examination. https://guidelines.euipo.europa.eu/binary/2058843/2000000000, 2023. Accessed: 24 August 2023.

[45] EUIPO. EU trade mark legal texts. https://euipo.europa.eu/ohimportal/en/eu-trade-mark-legal-texts, 2023. Accessed: 24 August 2023.

[46] EUIPO. TMview. https://www.tmdn.org/TMview/welcome#/tmview, 2023. Accessed: 28 October 2023.

[47] EUIPO. EU trade mark reform - summary of changes applying from 1 october 2017. https://euipo.europa.eu/tunnel-web/secure/webdav/guest/document_library/contentPdfs/law_and_practice/legal_reform/Overview_changes_en.pdf, 2023. Accessed: 24 August 2023.

[48] C.J. Fall and C. Giraud-Carrier. Searching trademark databases for verbal similarities. *World Patent Information*, 27(2):135–143, 2005. ISSN 0172-2190. doi: https://doi.org/10.1016/j.wpi.2004.12.002. URL https://www.sciencedirect.com/science/article/pii/S017221900400153X.

[49] N. Faller and P. Salenbauch. Plurix: a multiprocessing unix-like operating system. In *Proceedings of the Second Workshop on Workstation Operating Systems*, pages 29–36, 1989. doi: 10.1109/WWOS.1989.109264.

[50] Tom Fawcett. Introduction to roc analysis. *Pattern Recognition Letters*, 27:861–874, 06 2006. doi: 10.1016/j.patrec.2005.10.010.

[51] Yitong Feng, Cunzhao Shi, Chengzuo Qi, Jian Xu, Baihua Xiao, and Chunheng Wang. Aggregation of reversal invariant features from edge images for large-scale trademark retrieval. In *2018 4th International Conference on Control, Automation and Robotics (ICCAR)*, pages 384–388, 2018. doi: 10.1109/ICCAR.2018.8384705.

[52] Ilanah Fhima and Catrina Denvir. An empirical analysis of the likelihood of confusion factors in european trade mark law. *IIC-International Review of Intellectual Property and Competition Law*, 46(3):310–339, 2015.

[53] Carol Friedman and Robert Sideli. Tolerating spelling errors during patient validation. *Computers and Biomedical Research*, 25(5):486–509, 1992. ISSN 0010-4809. doi: https://doi.org/10.1016/0010-4809(92)90005-U. URL https://www.sciencedirect.com/science/article/pii/001048099290005U.

[54] Ryosuke Fujita and Takahiro Hayashi. Vector image retrieval based on approximation of bezier curves with line segments. In *Proceedings of 2011 IEEE Pacific Rim Conference on Communications, Computers and Signal Processing*, pages 431–436, 2011. doi: 10.1109/PACRIM.2011.6032932.

[55] Najlah Gali, Radu Mariescu-Istodor, Damien Hostettler, and Pasi Fränti. Framework for syntactic string similarity measures. *Expert Systems with Applications*, 129: 169–185, 2019. ISSN 0957-4174. doi: https://doi.org/10.1016/j.eswa.2019.03.048. URL https://www.sciencedirect.com/science/article/pii/S0957417419302222.

[56] Simon Geiregat. Trade mark protection for smells, tastes and feels critical analysis of three non-visual signs in the eu. *International Review of Intellectual Property and Competition Law*, 53:219–245, 2022. URL https://doi.org/10.1007/s40319-022-01160-3.

[57] German Patent and Trade Mark Office. Goods and Services. https://www.dpma.de/english/trade_marks/classification/goods_and_services/index.html, 2023. Accessed: 18 October 2023.

[58] Jan-Mark Geusebroek, Gertjan J Burghouts, and Arnold WM Smeulders. The amsterdam library of object images. *International Journal of Computer Vision*, 61: 103–112, 2005.

[59] Osamu Gotoh. An improved algorithm for matching biological sequences. *Journal of Molecular Biology*, 162(3):705–708, 1982. ISSN 0022-2836. doi: https://doi.org/10.1016/0022-2836(82)90398-9. URL https://www.sciencedirect.com/science/article/pii/0022283682903989.

[60] Anjali Goyal, Ekta Walia, and Harvinder Singh Saini. Improved accuracy in shape based image retrieval with complex zernike moments using wavelets. In *2009 2nd International Congress on Image and Signal Processing*, pages 1–5, 2009. doi: 10.1109/CISP.2009.5304118.

[61] Isabelle Guyon, Jason Weston, Stephen Barnhill, and Vladimir Vapnik. Gene selection for cancer classification using support vector machines. *Machine learning*, 46:389–422, 2002.

[62] Rishin Haldar and Debajyoti Mukhopadhyay. Levenshtein distance technique in dictionary lookup methods: An improved approach. *arXiv preprint arXiv:1101.1232*, 2011.

[63] Maximilian Haller. *TMSIM-500*. Harvard Dataverse, 2023. doi: 10.7910/DVN/PNFQLC. URL https://doi.org/10.7910/DVN/PNFQLC.

[64] R. W. Hamming. Error detecting and error correcting codes. *The Bell System Technical Journal*, 29(2):147–160, 1950. doi: 10.1002/j.1538-7305.1950.tb00463.x.

[65] Tin Kam Ho. Random decision forests. In *Proceedings of 3rd International Conference on Document Analysis and Recognition*, volume 1, pages 278–282 vol.1, 1995. doi: 10.1109/ICDAR.1995.598994.

[66] Tin Kam Ho. The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(8):832–844, 1998. doi: 10.1109/34.709601.

[67] Zhiling Hong and Qingshan Jiang. Hybrid content-based trademark retrieval using region and contour features. In *22nd International Conference on Advanced Information Networking and Applications-Workshops (aina workshops 2008)*, pages 1163–1168. IEEE, 2008.

[68] Daniel J. Howard, Roger A. Kerin, and Charles Gengler. The effects of brand name similarity on brand source confusion: Implications for trademark infringement. *Journal of Public Policy & Marketing*, 19(2):250–264, 2000. doi: 10.1509/jppm.19.2.250.17131. URL https://doi.org/10.1509/jppm.19.2.250.17131.

[69] Yu-Lin Hsu. The complexity and recognizable information business trademarks design applied fractal analysis and fractal dimension. In *2016 International Conference on Advanced Materials for Science and Engineering (ICAMSE)*, pages 408–411, 2016. doi: 10.1109/ICAMSE.2016.7840360.

[70] Intellectual Property Office of Vietnam. Trademark database. https://decisions.ch/. Accessed: 10 November 2023.

[71] Yacine Izza, Alexey Ignatiev, and João Marques-Silva. On explaining decision trees. *CoRR*, abs/2010.11034, 2020. URL https://arxiv.org/abs/2010.11034.

[72] Anul K. Jain and Aditya Vailaya. Shape-based retrieval: A case study with trademark image databases. *Pattern Recognition*, 31(9):1369–1390, 1998. ISSN 0031-3203. doi: https://doi.org/10.1016/S0031-3203(97)00131-3. URL https://www.sciencedirect.com/science/article/pii/S0031320397001313.

[73] Dayanand Jamkhandikar and V.D. Mytri. Css based trademark retrieval system. In *2014 International Conference on Electronic Systems, Signal Processing and Computing Technologies*, pages 129–133, 2014. doi: 10.1109/ICESC.2014.27.

[74] Matthew A. Jaro. Advances in record-linkage methodology as applied to matching the 1985 census of tampa, florida. *Journal of the American Statistical Association*, 84(406):414–420, 1989. doi: 10.1080/01621459.1989.10478785. URL https://www.tandfonline.com/doi/abs/10.1080/01621459.1989.10478785.

[75] Hui Jiang, Chong-Wah Ngo, and Hung-Khoon Tan. Gestalt-based feature similarity measure in trademark database. *Pattern recognition*, 39(5):988–1001, 2006.

[76] Yafei Jiang. Characterizing decades of technological advances with graph neural networks: An innovation network perspective. In *2021 International Conference on Digital Society and Intelligent Systems (DSInS)*, pages 128–133, 2021. doi: 10.1109/DSInS54396.2021.9670622.

[77] K. Kameyama, N. Oka, and K. Toraichi. Optimal parameter selection in image similarity evaluation algorithms using particle swarm optimization. In *2006 IEEE International Conference on Evolutionary Computation*, pages 1079–1086, 2006. doi: 10.1109/CEC.2006.1688429.

[78] K. Kasravi and M. Risov. Multivariate patent similarity detection. In *2009 42nd Hawaii International Conference on System Sciences*, pages 1–8, 2009. doi: 10.1109/HICSS.2009.318.

[79] T. Kato, T. Kurita, H. Shimogaki, T. Mizutori, and K. Fujimura. Cognitive view mechanism for multimedia database system. In *[1991] Proceedings. First International Workshop on Interoperability in Multidatabase Systems*, pages 179–186, 1991. doi: 10.1109/IMS.1991.153702.

[80] Hae-Kwang Kim, Jong-Deuk Kim, Dong-Gyu Sim, and Dae-Il Oh. A modified zernike moment shape descriptor invariant to translation, rotation and scale for similarity-based image retrieval. In *2000 IEEE International Conference on Multimedia and Expo. ICME2000. Proceedings. Latest Advances in the Fast Changing World of Multimedia (Cat. No.00TH8532)*, volume 1, pages 307–310 vol.1, 2000. doi: 10.1109/ICME.2000.869602.

[81] Barbara Kitchenham and Stuart Charters. Guidelines for performing systematic literature reviews in software engineering. Technical Report EBSE-2007-01, Software Engineering Group School of Computer Science and Mathematics at Keeele University and Department of Computer Science at University of Durham, 2017.

[82] M.M. Klee. The perils of picking a trademark. *IEEE Engineering in Medicine and Biology Magazine*, 17(5):140–, 1998. doi: 10.1109/51.715498.

[83] Grzegorz Kondrak. N-gram similarity and distance. In *SPIRE*, 2005. URL `https://api.semanticscholar.org/CorpusID:7481332`.

[84] Varlam Kutateladze. The kernel trick for nonlinear factor modeling. *International Journal of Forecasting*, 38(1):165–177, 2022.

[85] Nojun Kwak. Nonlinear projection trick in kernel methods: An alternative to the kernel trick. *IEEE Transactions on Neural Networks and Learning Systems*, 24(12): 2113–2119, 2013. doi: 10.1109/TNNLS.2013.2272292.

[86] P.W.H. Kwan, K. Kameyama, and K. Toraichi. Connecting image similarity retrieval with consistent labeling problem by introducing a match-all label. In *10th IEEE International Conference on Fuzzy Systems. (Cat. No.01CH37297)*, volume 3, pages 1384–1387 vol.2, 2001. doi: 10.1109/FUZZ.2001.1008916.

[87] P.W.H. Kwan, K. Kameyama, and K. Toraichi. Trademark retrieval by relaxation matching on fluency function approximated image contours. In *2001 IEEE Pacific Rim Conference on Communications, Computers and Signal Processing (IEEE Cat.*

No.01CH37233), volume 1, pages 255–258 vol.1, 2001. doi: 10.1109/PACRIM.2001.953571.

[88] P.W.H. Kwan, K. Toraichi, K. Kameyama, F. Kawazoe, and K. Nakamura. Tast-trademark application assistant. In *Proceedings. International Conference on Image Processing*, volume 1, pages I–I, 2002. doi: 10.1109/ICIP.2002.1038167.

[89] Wei-Wei Lai, Wing W. Y. Ng, Patrick P.K. Chan, and Daniel S. Yeung. Trademark classification by shape using ensemble of rbfnns. In *2010 International Conference on Machine Learning and Cybernetics*, volume 1, pages 391–396, 2010. doi: 10.1109/ICMLC.2010.5581030.

[90] Tian Lan, Xiaoyi Feng, Lei Li, and Zhaoqiang Xia. Similar trademark image retrieval based on convolutional neural network and constraint theory. In *2018 Eighth International Conference on Image Processing Theory, Tools and Applications (IPTA)*, pages 1–6, 2018. doi: 10.1109/IPTA.2018.8608162.

[91] Yann Lecun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L.D. Jackel. Handwritten digit recognition with a back-propagation network. In David Touretzky, editor, *Advances in Neural Information Processing Systems (NIPS 1989), Denver, CO*, volume 2, pages 396–404. Morgan Kaufmann, 1990.

[92] Sang-Mi Lee, John Haozhong Xin, and Stephen Westland. Evaluation of image similarity by histogram intersection. *Color Research & Application: Endorsed by Inter-Society Color Council, The Colour Group (Great Britain), Canadian Society for Color, Color Science Association of Japan, Dutch Society for the Study of Color, The Swedish Colour Centre Foundation, Colour Society of Australia, Centre Français de la Couleur*, 30(4):265–274, 2005.

[93] Mark A Lemley and Mark McKenna. Irrelevant confusion. *Stan. L. Rev.*, 62:413, 2009.

[94] Goh Wee Leng and D.P. Mital. A system for trademark pattern registration and recognition. In *1998 Second International Conference. Knowledge-Based Intelligent Electronic Systems. Proceedings KES'98 (Cat. No.98EX111)*, volume 3, pages 363–369 vol.3, 1998. doi: 10.1109/KES.1998.725995.

[95] Wing Ho Leung and Tsuhan Chen. Retrieval of sketches based on spatial relation between strokes. In *Proceedings. International Conference on Image Processing*, volume 1, pages I–I, 2002. doi: 10.1109/ICIP.2002.1038173.

[96] Wing Ho Leung and Tsuhan Chen. Retrieval of hand-drawn sketches with partial matching. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03).*, volume 3, pages III–5, 2003. doi: 10.1109/ICASSP.2003.1199093.

[97] Vladimir I Levenshtein et al. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pages 707–710. Soviet Union, 1966.

[98] Lei Li, Dongsheng Wang, and Guohua Cui. Trademark image retrieval using region zernike moments. In *2008 Second International Symposium on Intelligent Information Technology Application*, volume 2, pages 301–305, 2008. doi: 10.1109/IITA.2008.330.

[99] Daryl Lim. Trademark confusion simplified: A new framework for multifactor tests. *Berkeley Tech. LJ*, 37:867, 2022.

[100] Yingchi Liu, Quanzhi Li, Changlong Sun, and Luo Si. Similar trademark detection via semantic, phonetic and visual similarity information. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2025–2030, 2021.

[101] Minh-Thang Luong and Christopher D. Manning. Achieving open vocabulary neural machine translation with hybrid word-character models. *CoRR*, abs/1604.00788, 2016. URL `http://arxiv.org/abs/1604.00788`.

[102] Ilias Maglogiannis, Kostas Karpouzis, Manolis Wallace, and John Soldatos, editors. *Emerging Artificial Intelligence Applications in Computer Engineering - Real Word AI Systems with Applications in eHealth, HCI, Information Retrieval and Pervasive Technologies*, volume 160 of *Frontiers in Artificial Intelligence and Applications*, 2007. IOS Press. ISBN 978-1-58603-780-2.

[103] Stefan Martin and Jonathan Boyd. software: clear and precise? *Journal of Intellectual Property Law & Practice*, 16(6):459–461, 2021.

[104] Gonzalo Medina. How to construct a confusion matrix in LaTeX? `https://tex.stackexchange.com/a/20284`, 2011. Accessed: 25 October 2023.

[105] Thomas Heinz Meitinger. Verwechslungsgefahr. *Ohne Anwalt zur Marke: Anleitung zum Erwerb wertvoller Marken*, pages 45–59, 2021.

[106] George Miaoulis and Nancy d'Amato. Consumer confusion & trademark infringement: Presents a new, broadened concept of consumer confusion, illustrated by research results in the tic tac® case. *Journal of Marketing*, 42(2):48–55, 1978.

[107] Matthew Michelson and Craig Knoblock. Unsupervised information extraction from unstructured, ungrammatical data sources on the world wide web. *IJDAR*, 10:211–226, 12 2007. doi: 10.1007/s10032-007-0052-2.

[108] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space, 2013.

[109] John Mingers. An empirical comparison of pruning methods for decision tree induction. *Machine learning*, 4:227–243, 1989.

[110] Tom M Mitchell. *Machine Learning*. McGraw-Hill Professional, New York, 1997. ISBN 978-0070428072.

[111] Fatahiyah 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. doi: 10.1109/SMC.2014.6974151.

[112] V. Mounika, N. Raghavendra Sai, Vasantha Bhavani, and P S V S Sridhar. Interest flooding attack detection method in ndn networks. In *2021 2nd International Conference on Smart Electronics and Communication (ICOSEC)*, pages 298–307, 2021. doi: 10.1109/ICOSEC51865.2021.9591714.

[113] MPEG MPEG. Overview (version 8.0), july 2002. iso. Technical report, IEC JTC1/SC29/WG11, 7.

[114] Saul B. Needleman and Christian D. Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48(3):443–453, 1970. ISSN 0022-2836. doi: https://doi.org/10.1016/0022-2836(70)90057-4. URL `https://www.sciencedirect.com/science/article/pii/0022283670900574`.

[115] Mário Popolin Neto and Fernando V Paulovich. Explainable matrix-visualization for global and local interpretability of random forest classification ensembles. *IEEE Transactions on Visualization and Computer Graphics*, 27(2):1427–1437, 2020.

[116] Tien Dung Nguyen, Huu Hiep Hai Nguyen, and Thanh Ha Le. Trademark image retrieval based on scale, rotation, translation invariant features. In *The 2013 RIVF International Conference on Computing Communication Technologies - Research, Innovation, and Vision for Future (RIVF)*, pages 282–285, 2013. doi: 10.1109/RIVF.2013.6719908.

[117] Steven John Olsen. Mixed signals in trademark's likelihood of confusion law: Does quality matter. *Val. UL Rev.*, 44:659, 2009.

[118] Xing Liang Lazaros Toumanidis Georgia Sakellari Charalampos Patrikakis George Loukas Panagiotis Kasnesis, Ryan Heartfield. Transformer-based identification of stochastic information cascades in social networks using text and image similarity. In *Journal of Applied Soft Computing*, 2021.

[119] Kitsuchart Pasupa and Wisuwat Sunhem. A comparison between shallow and deep architecture classifiers on small dataset. In *2016 8th International Conference on Information Technology and Electrical Engineering (ICITEE)*, pages 1–6, 2016. doi: 10.1109/ICITEED.2016.7863293.

[120] Geoffroy Peeters and Emmanuel Deruty. Sound indexing using morphological description. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(3): 675–687, 2010. doi: 10.1109/TASL.2009.2038809.

[121] Jeffrey Pennington, Richard Socher, and Christopher Manning. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, October 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1162. URL https://aclanthology.org/D14-1162.

[122] Claudio A. Perez, Pablo A Estévez, Francisco J. Galdames, Daniel A. Schulz, Juan P. Perez, Diego Bastías, and Daniel R. Vilar. Trademark image retrieval using a combination of deep convolutional neural networks. In *2018 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7, 2018. doi: 10.1109/IJCNN.2018.8489045.

[123] Lawrence Philips. The double metaphone search algorithm. *C/C++ Users J.*, 18 (6):3843, jun 2000. ISSN 1075-2838.

[124] Lawrence Philips. Metaphone 3 v2.1.3. https://github.com/OpenRefine/OpenRefine/blob/master/main/src/com/google/refine/clustering/binning/Metaphone3.java, 2010. Accessed: 7 November 2023.

[125] Latika Pinjarkar, Manisha Sharma, and Smita Selot. Improved trademark image retrieval system using relevance feedback. In *2016 8th International Conference on Computational Intelligence and Communication Networks (CICN)*, pages 298–304, 2016. doi: 10.1109/CICN.2016.65.

[126] Ya-Li Qi. A relevance feedback retrieval method based on tamura texture. In *2009 Second International Symposium on Knowledge Acquisition and Modeling*, volume 3, pages 174–177, 2009. doi: 10.1109/KAM.2009.39.

[127] Ya-Li Qi. A relevance feedback method to trademark retrieval based on svm. In *2009 International Symposium on Computer Network and Multimedia Technology*, pages 1–4, 2009. doi: 10.1109/CNMT.2009.5374552.

[128] Zhipeng Qiu and Zheng Wang. Technology forecasting based on semantic and citation analysis of patents: A case of robotics domain. *IEEE Transactions on Engineering Management*, 69(4):1216–1236, 2022. doi: 10.1109/TEM.2020.2978849.

[129] S. Ravela and R. Manmatha. On computing global similarity in images. In *Proceedings Fourth IEEE Workshop on Applications of Computer Vision. WACV'98 (Cat. No.98EX201)*, pages 82–87, 1998. doi: 10.1109/ACV.1998.732862.

[130] Steffen Reinhard. *Gegenstand und Prüfungsmaßstab der markenrechtlichen Verwechslungsgefahr*, volume 150. Mohr Siebeck, 2020.

[131] Mark D Robins. Actual confusion in trademark infringement litigation: Restraining subjectivity through a factor-based approach to valuing evidence. *Nw. J. Tech. & Intell. Prop.*, 2:117, 2003.

[132] Eleonora Rosati. The absolute ground for refusal or invalidity in article 7(1)(e)(iii) eutmr/4(1)(e)(iii) eutmd: in search of the exclusions own substantial value. *Journal of Intellectual Property Law*, pages 1–20, 2020.

[133] Eleonora Rosati. Tell me what you c: Chanel loses monogram battle against huawei. *Journal of Intellectual Property Law & Practice*, 16(6):458–469, 2021.

[134] Marçal Rusiñol, Farshad Noorbakhsh, Dimosthenis Karatzas, Ernest Valveny, and Josep Lladós. Perceptual image retrieval by adding color information to the shape context descriptor. In *2010 20th International Conference on Pattern Recognition*, pages 1594–1597, 2010. doi: 10.1109/ICPR.2010.394.

[135] Mark Schweizer. Schweizerische Mitteilungen über Immaterialgüterrecht 1951-1996. `https://decisions.ch/`. Accessed: 10 November 2023.

[136] scikit-learn developers. Kernel functions, scikit-learn v1.3.2. `https://scikit-learn.org/stable/modules/svm.html#svm-kernels`, 2023. Accessed: 24 October 2023.

[137] G Thamarai Selvi, Amal Hamdy, K Sri Vijaya, Nellore Manoj Kumar, L Pallavi, and Kanusu Srinivasa Rao. Reliable and efficient image processing and deep machine learning for large-scale digital image retrieval. In *2022 International Interdisciplinary Humanitarian Conference for Sustainability (IIHC)*, pages 1481–1484, 2022. doi: 10.1109/IIHC55949.2022.10059657.

[138] Rossitza Setchi and Fatahiyah Mohd Anuar. Multi-faceted assessment of trademark similarity. *Expert Systems with Applications*, 65:16–27, 2016. ISSN 0957-4174. doi: https://doi.org/10.1016/j.eswa.2016.08.028. URL `https://www.sciencedirect.com/science/article/pii/S0957417416304213`.

[139] Rossitza Setchi and Fatahiyah Mohd Anuar. Multi-faceted assessment of trademark similarity. *Expert Systems with Applications*, 65:16–27, 2016.

[140] C. E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423, 1948. doi: 10.1002/j.1538-7305.1948.tb01338.x.

[141] Day-Fann Shen, Li Jin, H.T. Chang, and H.H.P. Wu. Tademark retrieval based on block feature index code. In *IEEE International Conference on Image Processing 2005*, volume 3, pages III–177, 2005. doi: 10.1109/ICIP.2005.1530357.

[142] Shohei Shirataki and Saneyasu Yamaguchi. A study on interpretability of decision of machine learning. In *2017 IEEE International Conference on Big Data (Big Data)*, pages 4830–4831, 2017. doi: 10.1109/BigData.2017.8258557.

[143] M. S. Shirdhonkar and Manesh B. Kokare. Document image retrieval using signature as query. In *2011 2nd International Conference on Computer and Communication Technology (ICCCT-2011)*, pages 66–70, 2011. doi: 10.1109/ICCCT.2011.6075200.

[144] T.F. Smith and M.S. Waterman. Identification of common molecular subsequences. *Journal of Molecular Biology*, 147(1):195–197, 1981. ISSN 0022-2836. doi: https://doi.org/10.1016/0022-2836(81)90087-5. URL https://www.sciencedirect.com/science/article/pii/0022283681900875.

[145] Thorvald Sorensen. A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on danish commons. *Biologiske skrifter*, 5:1–34, 1948.

[146] Irina Suslina and Polina Mineeva. Use of digital technologies for optimizing the handling of trademark applications. *Procedia Computer Science*, 169:435–439, 2020. ISSN 1877-0509. doi: https://doi.org/10.1016/j.procs.2020.02.242. URL https://www.sciencedirect.com/science/article/pii/S1877050920303665. Postproceedings of the 10th Annual International Conference on Biologically Inspired Cognitive Architectures, BICA 2019 (Tenth Annual Meeting of the BICA Society), held August 15-19, 2019 in Seattle, Washington, USA.

[147] Charles V Trappey, Amy JC Trappey, and Sam C-C Lin. Intelligent trademark similarity analysis of image, spelling, and phonetic features using machine learning methodologies. *Advanced Engineering Informatics*, 45:101120, 2020.

[148] Priyanka Tripathi and Ajay Kumar Indoria. Extraction and recognition of multi-oriented text from trademark images. In *2015 International Conference on Cognitive Computing and Information Processing(CCIP)*, pages 1–5, 2015. doi: 10.1109/CCIP.2015.7100700.

[149] Osman Tursun, Cemal Aker, and Sinan Kalkan. A large-scale dataset and benchmark for similar trademark retrieval. *CoRR*, abs/1701.05766, 2017. URL http://arxiv.org/abs/1701.05766.

[150] Shashank Upadhye. Trademark surveys: Identifying the relevant universe of confused consumers. *Fordham Intell. Prop. Media & Ent. LJ*, 8:549, 1997.

[151] Arjeton Uzairi, Arianit Kurti, and Zenun Kastrati. A deep learning-based solution for identification of figurative elements in trademark images. In *2023 4th International Conference on Computing, Mathematics and Engineering Technologies (iCoMET)*, pages 1–7, 2023. doi: 10.1109/iCoMET57998.2023.10099183.

[152] A. Vailaya, Yu Zhong, and A.K. Jain. A hierarchical system for efficient image retrieval. In *Proceedings of 13th International Conference on Pattern Recognition*, volume 3, pages 356–360 vol.3, 1996. doi: 10.1109/ICPR.1996.546970.

[153] Mark PJ Van der Loo et al. The stringdist package for approximate string matching. *R J.*, 6(1):111, 2014.

[154] Valeriy S Vykhovanets, J Du, and SA Sakulin. An overview of phonetic encoding algorithms. *Automation and Remote Control*, 81:1896–1910, 2020.

[155] Bin Wang, Angela Wang, Fenxiao Chen, Yuncheng Wang, and C.-C. Jay Kuo. Evaluating word embedding models: methods and experimental results. *APSIPA Transactions on Signal and Information Processing*, 8:e19, 2019. doi: 10.1017/ATSIP.2019.12.

[156] Lipo Wang. *Support Vector Machines: Theory and Applications (Studies in Fuzziness and Soft Computing)*. Springer-Verlag, Berlin, Heidelberg, 2005. ISBN 3540243887.

[157] Tianlei Wang, Zeliang Li, Ying Xu, Jiacong Chen, Angelo Genovese, Vincenzo Piuri, and Fabio Scotti. Few-shot steel surface defect recognition via self-supervised teacherstudent model with minmax instances similarity. *IEEE Transactions on Instrumentation and Measurement*, 72:1–16, 2023. doi: 10.1109/TIM.2023.3315404.

[158] Yong-jiao Wang and Chun-feng Zheng. Trademark image retrieval based on shape and key local color features. In *2009 Second International Conference on Information and Computing Science*, volume 2, pages 325–328, 2009. doi: 10.1109/ICIC.2009.193.

[159] Shiuh-Ku Weng, Cliung-Ming Kuo, and Ta-Wen Kuan. New shape descriptor for trademarks retrieval. In *NSIP 2005. Abstracts. IEEE-Eurasip Nonlinear Signal and Image Processing, 2005.*, pages 48–, 2005. doi: 10.1109/NSIP.2005.1502312.

[160] William Winkler. String comparator metrics and enhanced decision rules in the fellegi-sunter model of record linkage. *Proceedings of the Section on Survey Research Methods*, 01 1990.

[161] M. Wolfe. Supercompilers, the amd opteron, and your cell phone. In *18th International Parallel and Distributed Processing Symposium, 2004. Proceedings.*, pages 98–, 2004. doi: 10.1109/IPDPS.2004.1303044.

[162] Jian-Kang Wu. Content-based indexing of multimedia databases. *IEEE Transactions on Knowledge and Data Engineering*, 9(6):978–989, 1997. doi: 10.1109/69.649320.

[163] Zhibiao Wu and Martha Palmer. Verbs semantics and lexical selection. pages 133–138, 01 1994. doi: 10.3115/981732.981751.

[164] Yijun Yan, Jinchang Ren, Yinsheng Li, James Windmill, and Winifred Ijomah. Fusion of dominant colour and spatial layout features for effective image retrieval of coloured logos and trademarks. In *2015 IEEE International Conference on Multimedia Big Data*, pages 306–311, 2015. doi: 10.1109/BigMM.2015.43.

[165] Wei Yu, Kuiyuan Yang, Hongxun Yao, Xiaoshuai Sun, and Pengfei Xu. Exploiting the complementary strengths of multi-layer cnn features for image retrieval. *Neurocomput.*, 237(C):235241, may 2017. ISSN 0925-2312. doi: 10.1016/j.neucom.2016.12.002. URL https://doi.org/10.1016/j.neucom.2016.12.002.

[166] Li Yujian and Liu Bo. A normalized levenshtein distance metric. *IEEE transactions on pattern analysis and machine intelligence*, 29(6):1091–1095, 2007.

[167] Siderite Zackwehdex. Super fast and accurate string distance algorithm: SIFT4. https://siderite.dev/blog/super-fast-and-accurate-string-distance.html, 2014. Accessed: 7 November 2023.

[168] Ahmed Zeggari, Fella Hachouf, and Sebti Foufou. Trademarks recognition based on local regions similarities. In *10th International Conference on Information Science, Signal Processing and their Applications (ISSPA 2010)*, pages 37–40, 2010. doi: 10.1109/ISSPA.2010.5605559.

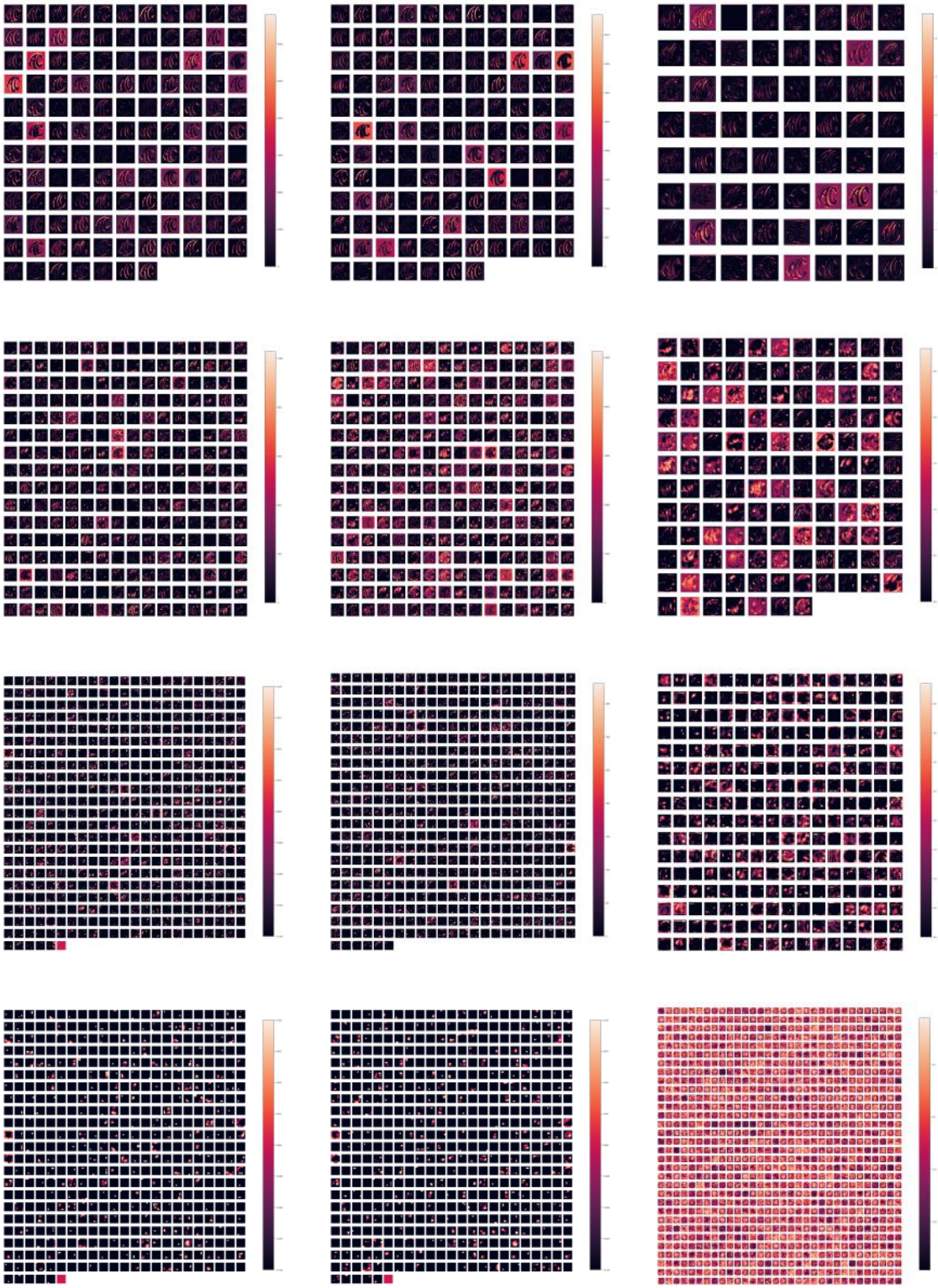
[169] Zhengze Zhou and Giles Hooker. Unbiased measurement of feature importance in tree-based methods. *ACM Trans. Knowl. Discov. Data*, 15(2), jan 2021. ISSN 1556-4681. doi: 10.1145/3429445. URL https://doi.org/10.1145/3429445.

[170] Bei-ji Zou and Marie Providence Umugwaneza. Shape-based trademark retrieval using cosine distance method. In *2008 Eighth International Conference on Intelligent Systems Design and Applications*, volume 2, pages 498–504, 2008. doi: 10.1109/ISDA.2008.161.

Appendix

Appendix A: CNN Activations

The images below show the activations for the VGG16, the VGG19 and the ResNet50 model (from left to right) using the image "002173717_contested.png" from the TMSIM-500 dataset. The first row shows the input for each model. Each following row shows the output of a convolutional block.



Appendix B: Literature Review Protocol

This protocol is part of the systematic literature review performed to find existing methods for computing trademark similarities so as to answer RQ1. For the sake of reproducibility, the search operations and the selection of literature is documented in this protocol according to the guidelines defined by Kitchenham and Charters [81].

This literature review only considers studies that compare similarity measures on trademark data or introduce new methods to measure the similarity of word or figurative marks. Trademark data is not restricted to European trademark law. However, studies need to be in German or English and must be linked to a digital object identifier in order to be considered in this literature review. Furthermore, results are excluded from the review if they are not complete papers, like published abstracts.

Table A1 lists the queries and the respective search results. Search results that are in line with the inclusion criteria are listed in tables A2, A3, and A4. Search results excluded from the literature review are shown in table A5 with a note of why the result is excluded.

Tables A2, A3, and A4 contain the similarity aspect related to the problem addressed in the work, the proposed method, the related task, and the data used. In these tables, V, A, and C refer to visual, aural and conceptual similarity. The aspect "global" refers to the global assessment.

Search Engine	Query	Search Results	Excluded
IEEE Xplore	trademark similarity	[9], [116], [2], [51], [122], [30], [111], [125], [73], [170], [28], [25], [89], [168], [141], [98], [164], [94], [88], [87], [90], [32], [6], [80], [86], [129], [127], [54], [5], [126], [23], [95], [134], [162], [34], [77], [152], [60], [137]	[1], [69], [26], [151], [82], [148], [78], [157], [76], [38], [96], [79], [143], [120], [24], [49], [20], [161], [112]
IEEE Xplore	trademark phonetic similarity	[111]	
IEEE Xplore	trademark conceptual similarity	[9], [122], [151]	
IEEE Xplore	trademark semantic similarity	[9], [127], [126], [23], [128]	
IEEE Xplore	trademark string similarity	[9]	
ScienceDirect	trademark string similarity	[147], [138], [8], [48]	[146]
ScienceDirect	trademark conceptual similarity	[8]	
ScienceDirect	conceptual similarity of goods and services		
ScienceDirect	trademark phonetic similarity	[147], [138]	[48]
Google Scholar	trademark similarity	[32], [9], [147], [8]	[68], [4], [100], [75], [9], [48]

Table A1: Queries Performed and Search Results

Search Result	Year	Aspect	Proposed Method	Task	Data
Agrawal et al. [2]	2013	V	Zernike Moment + Curvelet Transform	Retrieval	NA
Alshowaish et al. [4]	2022	V	AlexNet, VGG16, ResNet-50	Retrieval	[149]
Alwis and Austin [6]	1998	V	Gestalt Features	Retrieval	NA
Alwis and Austin [5]	1999	V	Gestalt Features + CMM Neural Network	Retrieval	NA
Anuar et al. [8]	2013	C	WordNet	Retrieval	[135]
Anuar et al. [9]	2016	C	WordNet Ontology	Retrieval	[135]
Chalechale and Faramarzi [23]	2010	V	Edge Pixel Neighbouring Histogram + Histogram of Edge Directions	Retrieval	NA
Chen et al. [25]	2007	V	Region Orientation Information Entropy	Retrieval	NA
Ciocca and Schettini [28]	1999	V	Invariant Moments + Canny Edge Histogram + Wavelet Transform	Retrieval	[72]
Crysdian [30]	2014	V	Watershed Transform	Retrieval	NA
Eakins et al. [32]	1998	V	Boundary Shape Vector + Family Characteristics Vector	Retrieval	NA
Easttom [34]	2021	V	Structural Similarity Index	Classification	NA
Fall and Giraud-Carrier [48]	2005	A	Levenshtein, N-Gram, Damerau-Levenshtein + SoundEx, Double Metaphone, Editex	NA	[135]
Feng et al. [51]	2018	V	Modified SIFT	Retrieval	[149]
Fujita and Hayashi [54]	2011	V	Bezier Curves Approximation using Line Segments	Retrieval	NA
Goyal et al. [60]	2009	V	Zernike Moments + Wavelet Descriptors	Retrieval	NA
Jamkhandikar and Mytri [73]	2014	V	Curvature Scale Space Matching	Retrieval	[72]

Table A2: Included Search Results

Search Result	Year	Aspect	Proposed Method	Task	Data
Jiang et al. [75]	2006	V	Gestalt-based Features	Retrieval	NA
Kameyama et al. [77]	2006	V	Relaxation Matching + Particle Swarm Optimization	Retrieval	NA
Kim et al. [80]	2000	V	Modified Zernike Moment Descriptor	Retrieval	NA
Kwan et al. [86]	2001	V	Relaxation Labeling	Retrieval	NA
Kwan et al. [87]	2001	V	Multi-Stage Joint Points Extraction Algorithm + Relaxation Matching	Retrieval	NA
Kwan et al. [88]	2002	V	Multi-Stage Joint Points Extraction Algorithm + Relaxation Matching	Retrieval	NA
Lai et al. [89]	2010	V	Invariant Moments + RBFNN	Retrieval	NA
Lan et al. [90]	2018	V	Siamese VGG-F, Triplet Model	Retrieval	[149]
Leng and Mital [94]	1998	V	Invariant Moments + Color Histogram + Image-Background-Ratio	Retrieval	NA
Leung and Chen [95]	2002	V	Spatial Relation between Strokes	Retrieval	NA
Li et al. [98]	2008	V	Region Zernike Moments	Retrieval	NA
Mohd Anuar et al. [111]	2014	A	ALINE Algorithm	Retrieval	[135]
Liu et al. [100]	2021	V+A+C	Bi-Directional GRU with Character Embeddings, Phonetic Embeddings, Visual Embeddings, and Word Segmentation Encoding Position	Retrieval	NA
Nguyen et al. [116]	2013	V	RBRC Algorithm	Retrieval	[70]
Perez et al. [122]	2018	V	AlexNet, VGGNet, GoogLeNet	Retrieval	[149]
Pinjarkar et al. [125]	2016	V	Fourier Descriptor + Color Histrogram, Color Moments, Color Correlogram + Gabor Wavelet, Haar Wavelet	Retrieval	NA
Qi [126]	2009	V	Tamura Texture + SVM	Retrieval	NA

Table A3: Included Search Results

Search Result	Year	Aspect	Proposed Method	Task	Data
Qi [127]	2009	V	Tamura Texture + SVM	Retrieval	NA
Ravela and Manmatha [129]	1998	V	Curvature + Phase Histograms	Retrieval	NA
Rusiñol et al. [134]	2010	V	Shape Context Descriptor + Color Naming + Local Color Names Histogram	Retrieval	[58]
Selvi et al. [137]	2022	V	R2D2, SIFT	Retrieval	NA
Setchi and Anuar [138]	2016	Global	Multi-Faceted Assessment using Fuzzy Logic	Classification	[135]
Shen et al. [141]	2005	V	Block Feature Index Code	Retrieval	[113]
Trappey et al. [147]	2020	V+A	VGGnet + LCS + Cosine + word2vec + SoundEx, Metaphone, Dbmetaphone, NYSIIS	Classification	NA
Vailaya et al. [152]	1996	V	Canny Edge Histogram + Invariant Moments + Deformable Template Matching	Retrieval	NA
Wang and Zheng [158]	2009	V	Improved Wavelet Modulus Maxima + Key Local Color	Retrieval	NA
Wu [162]	1997	V	Learning based on Experiences and Perspectives	Retrieval	NA
Yan et al. [164]	2015	V	Linear Block Algorithm + Dominant Color Descriptor + Spatial Descriptor	Retrieval	NA
Zeggari et al. [168]	2010	V	Invariant Moments + Color Histogram	Retrieval	NA
Zou and Umugwaneza [170]	2008	V	Invariant Moments and Eccentricity + Entropy Histogram, Distance Histogram + Euclidean, Cosine Similarity	Retrieval	NA

Table A4: Included Search Results

Search Result	Reason for Exclusion
Agarwal et al. [1]	Digital object identifier missing
Buehler [20]	Does not address methods for trademark similarity computation
Charles [24]	Does not address methods for trademark similarity computation
Chen [26]	Mainly addresses trademark detection
Erol and Kossentini [38]	Addresses video objects
Faller and Salenbauch [49]	Does not address methods for trademark similarity computation
Howard et al. [68]	Does not address methods for trademark similarity computation
Hsu [69]	Does not address trademark similarity but trademark design
Jiang [76]	Does not address methods for trademark similarity computation
Kasravi and Risov [78]	Addresses patents instead of trademarks
Kato et al. [79]	Does not address methods for trademark similarity computation
Klee [82]	Does not address methods for trademark similarity computation
Leung and Chen [96]	Addresses partial matching but trademark retrieval requires whole matching
Mounika et al. [112]	Does not address methods for trademark similarity computation
Peeters and Deruty [120]	Does not address methods for comparing word or figurative marks
Qiu and Wang [128]	Does not address methods for trademark similarity computation
Shirdhonkar and Kokare [143]	Does not address methods for trademark similarity computation
Suslina and Mineeva [146]	Does not address methods for trademark similarity computation
Tripathi and Indoria [148]	Addesses text extraction from trademark images
Uzairi et al. [151]	Addresses labeling of trademark images with Vienna codes from the Vienna Classification
Wang et al. [157]	Addresses the detection of steel surface defects
Weng et al. [159]	Publication only contains abstract
Wolfe [161]	Does not address methods for trademark similarity computation

Table A5: Excluded Search Results

Appendix C: ROC Curves and Recall-Precision Curves

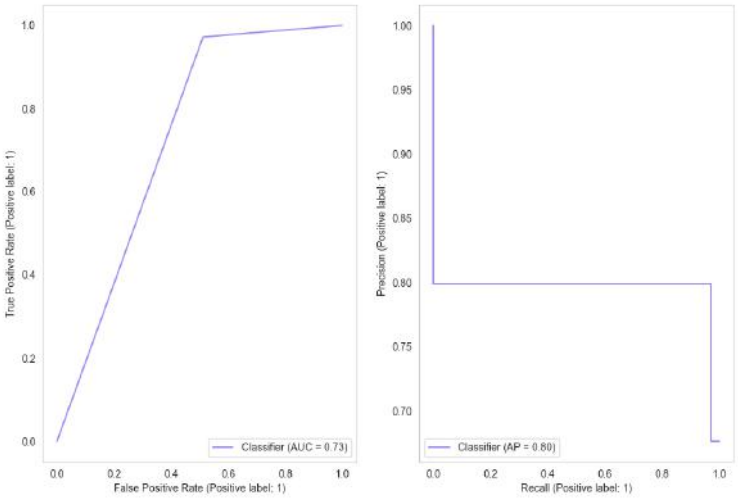


Figure A1: ROC Curve and Recall-Precision Curve: RF for Word Marks

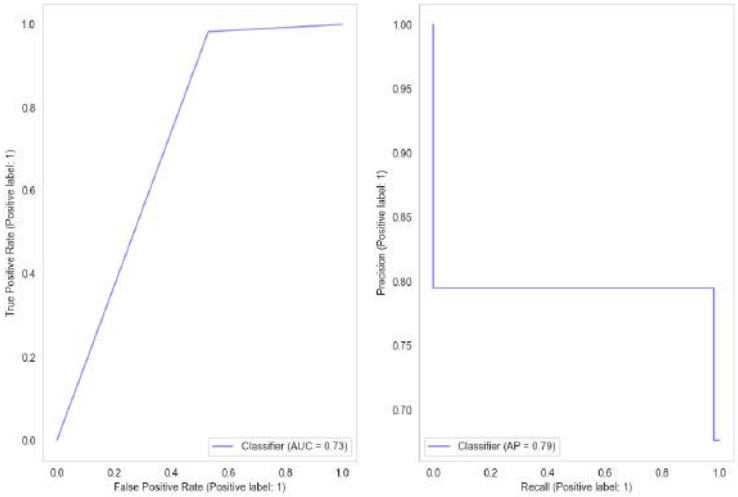


Figure A2: ROC Curve and Recall-Precision Curve: SVM for Word Marks

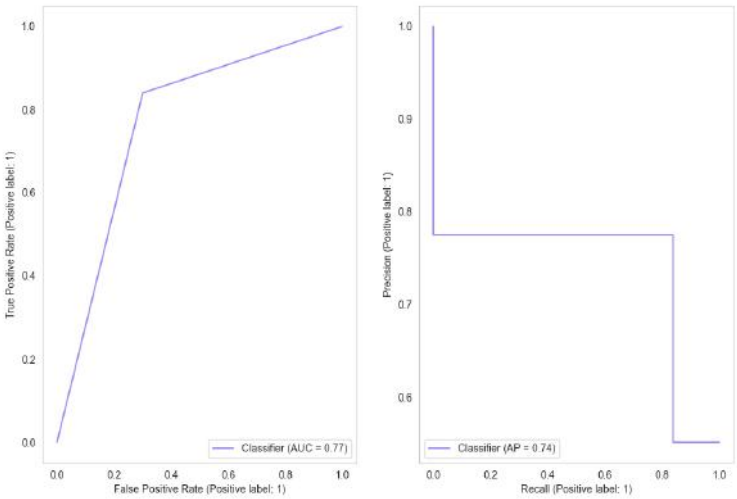


Figure A3: ROC Curve and Recall-Precision Curve: RF for Figurative Marks

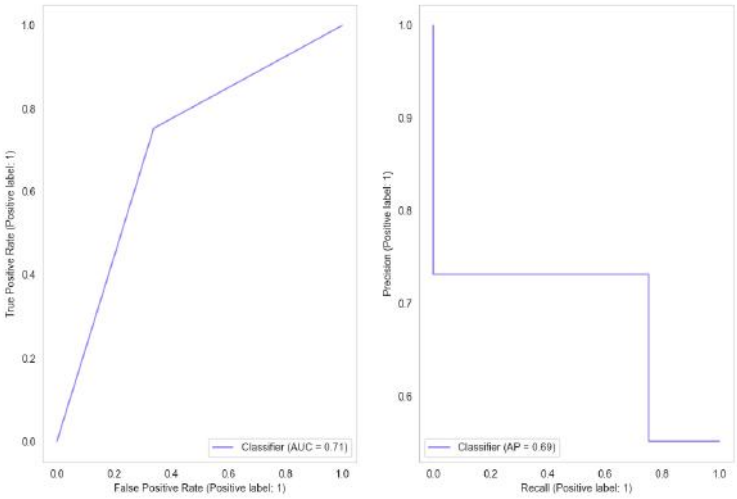


Figure A4: ROC Curve and Recall-Precision Curve: SVM for Figurative Marks

Appendix D: Learning Curves

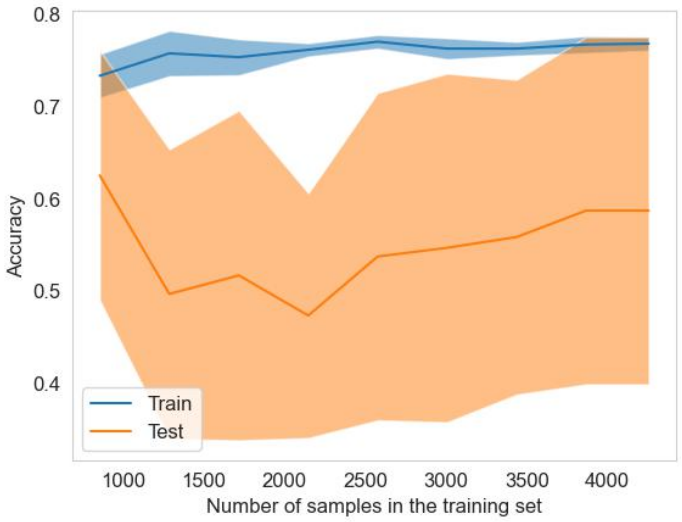


Figure A5: Learning Curves: RF for Word Marks

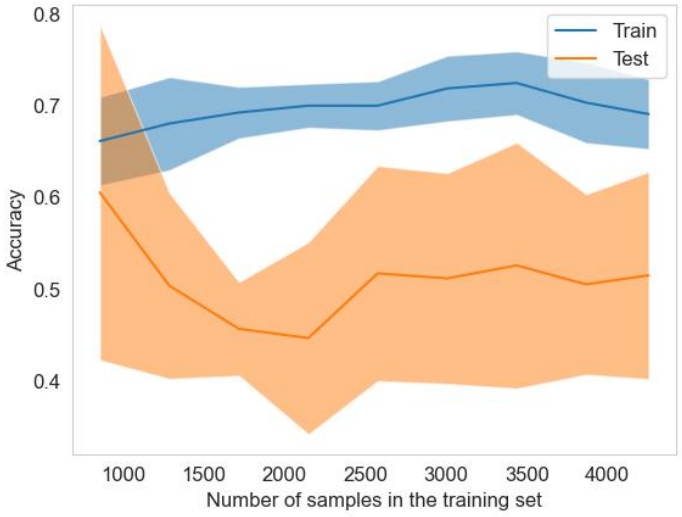


Figure A6: Learning Curves: SVM for Word Marks

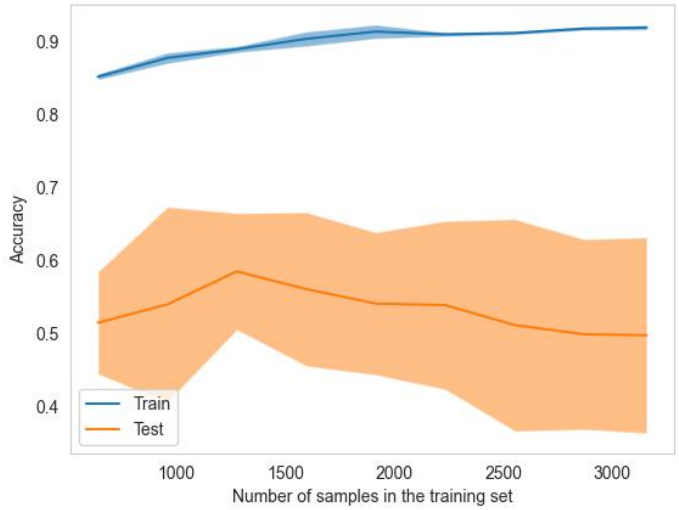


Figure A7: Learning Curves: RF for Figurative Marks

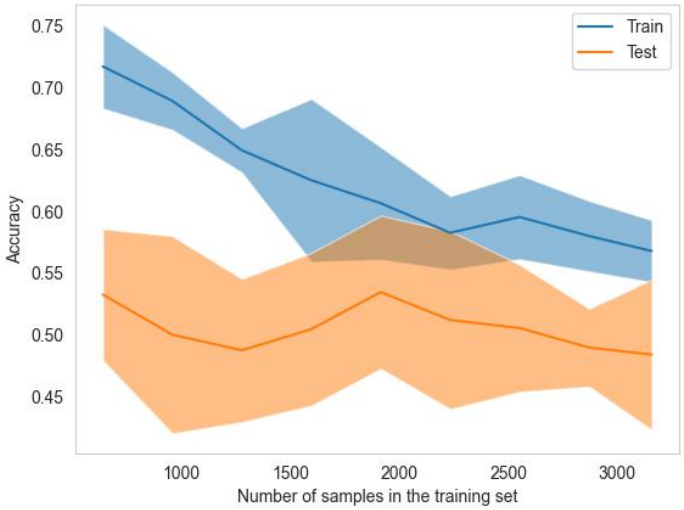


Figure A8: Learning Curves: SVM for Figurative Marks

Appendix E: SVM Decision Boundaries for Word Mark Data

Figures A9, A10, and A11 show the decision boundaries for the optimized SVM model for word mark data. Since this model operates on three features, the hyperplane cannot be visualized in one two-dimensional plot. Therefore, each figure shows the decision boundary for two variables only.

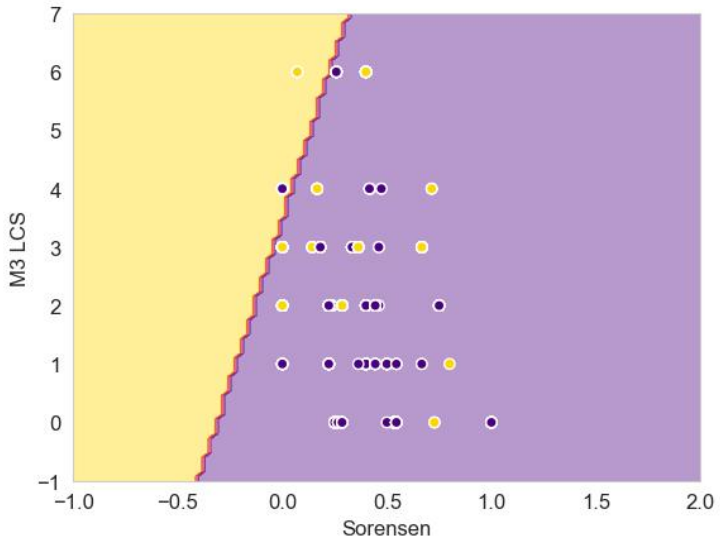


Figure A9: SVM for Word Mark Data: Decision Boundary for Visual Similarity and Aural Similarity

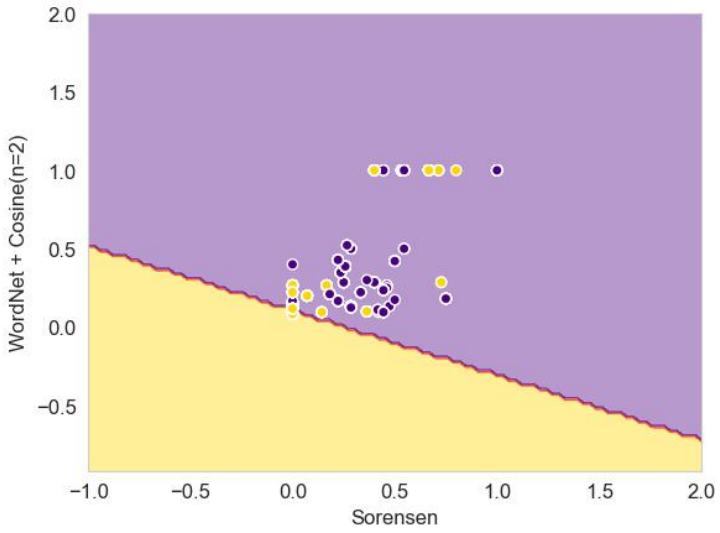


Figure A10: SVM for Word Mark Data: Decision Boundary for Visual Similarity and Conceptual Similarity

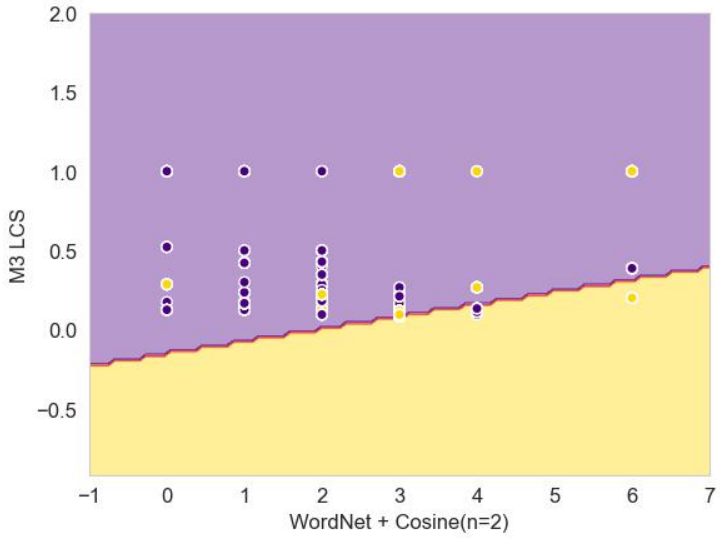


Figure A11: SVM for Word Mark Data: Decision Boundary for Aural Similarity and Conceptual Similarity

Appendix F: SVM Decision Boundaries for Figurative Mark Data

Figures A12, A13, A14, A15, A16, and A17 show the decision boundaries for the optimized SVM model for figurative mark data.

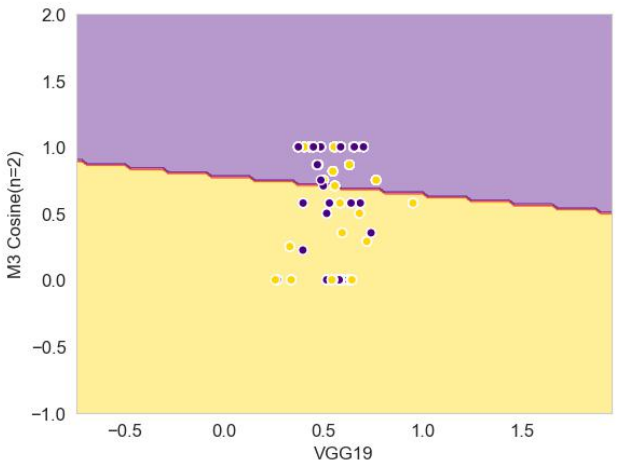


Figure A12: SVM for Figurative Mark Data: Decision Boundary for Visual Similarity and Aural Similarity

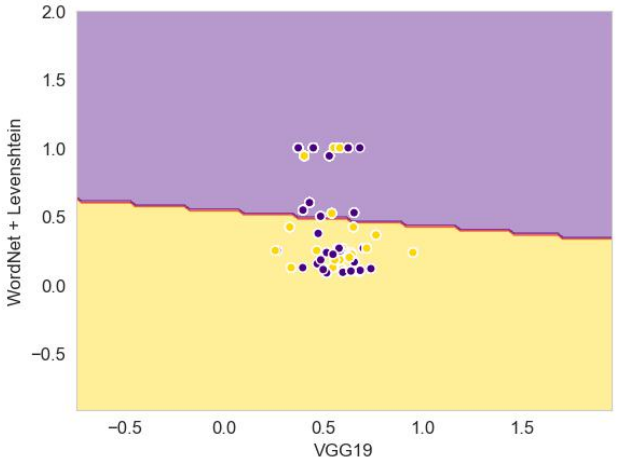


Figure A13: SVM for Figurative Mark Data: Decision Boundary for Visual Similarity and Conceptual Similarity

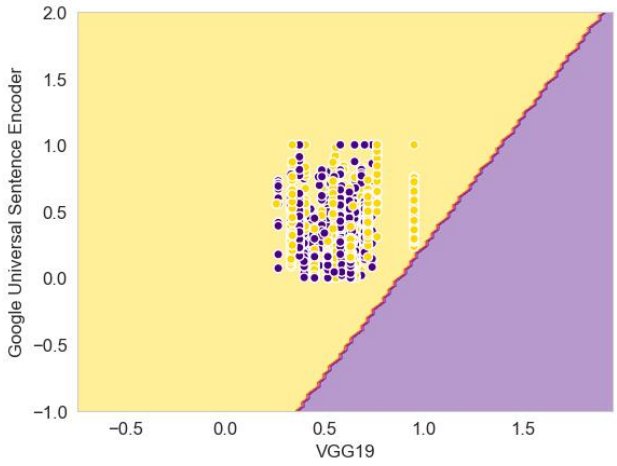


Figure A14: SVM for Figurative Mark Data: Decision Boundary for Visual Similarity and Item Similarity

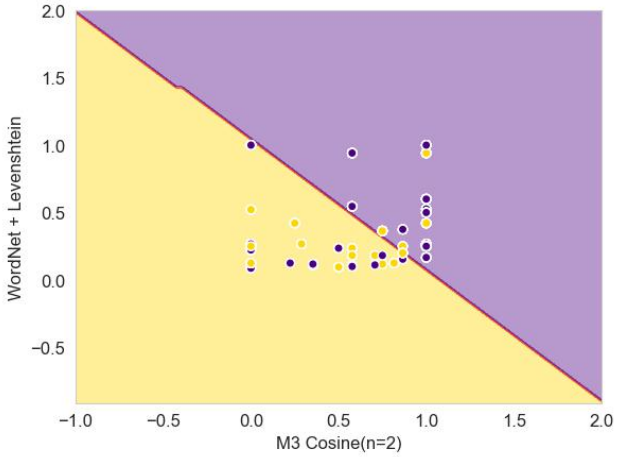


Figure A15: SVM for Figurative Mark Data: Decision Boundary for Aural Similarity and Conceptual Similarity

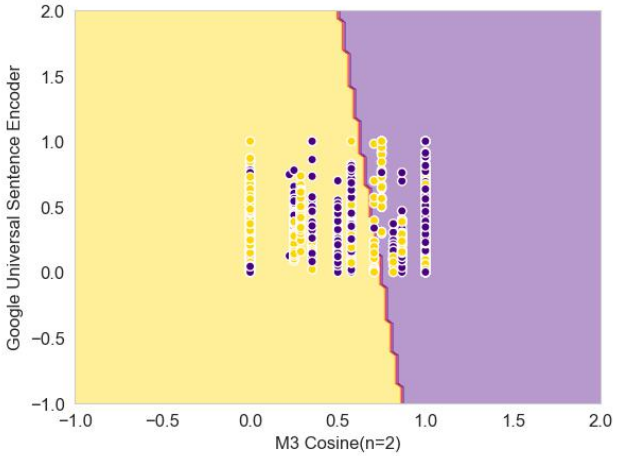


Figure A16: SVM for Figurative Mark Data: Decision Boundary for Aural Similarity and Item Similarity

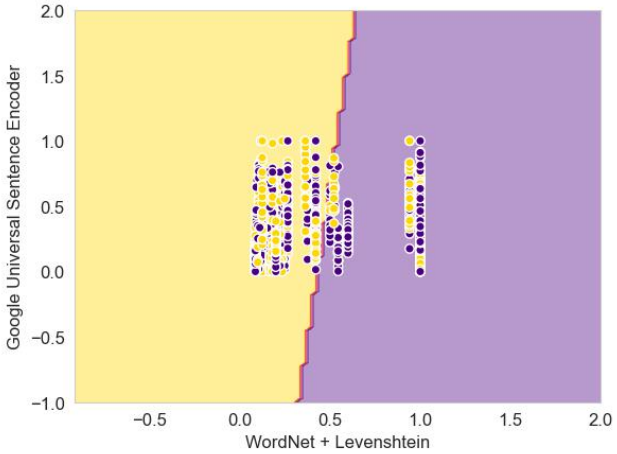


Figure A17: SVM for Figurative Mark Data: Decision Boundary for Conceptual Similarity and Item Similarity

Appendix G: Feature Performances

This section contains plots that compare the average F1-score achieved by every feature within a feature group. For SVMs, this comparison also includes scaling methods.

RF on Word Mark Data

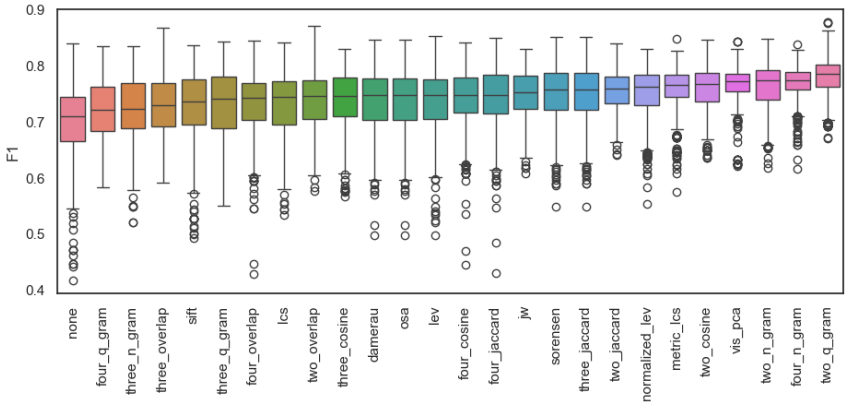


Figure A18: RF on Word Mark Data: Average Performance of Features for Visual Similarity

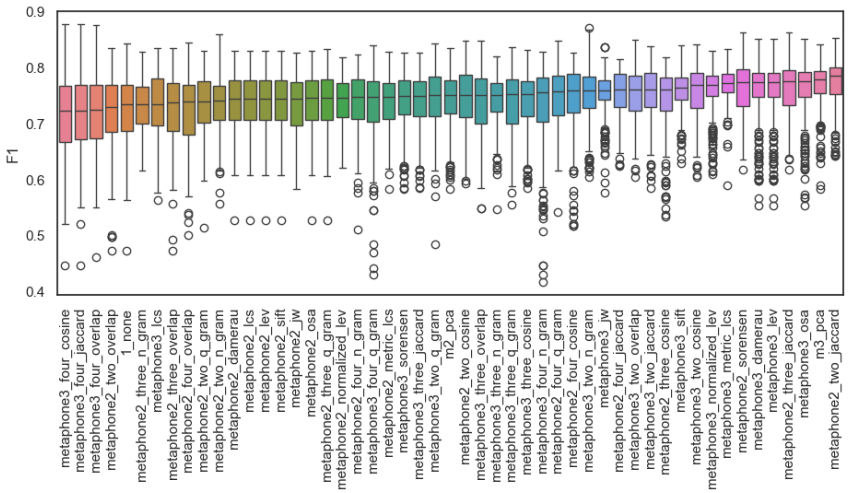


Figure A19: RF on Word Mark Data: Average Performance of Features for Aural Similarity

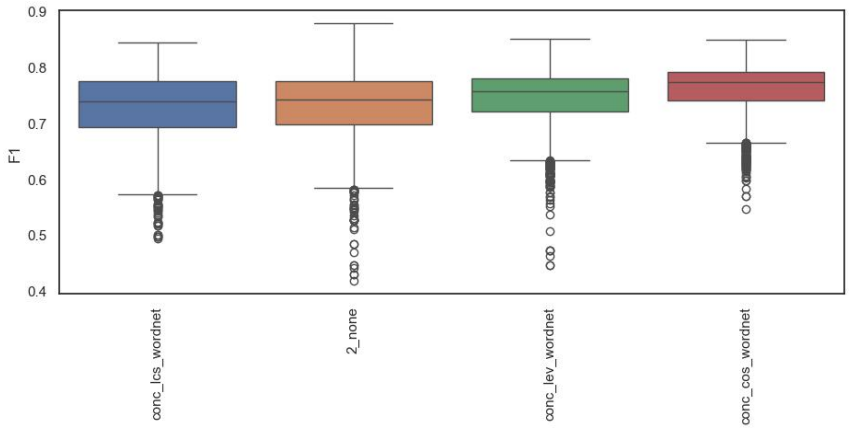


Figure A20: RF on Word Mark Data: Average Performance of Features for Conceptual Similarity

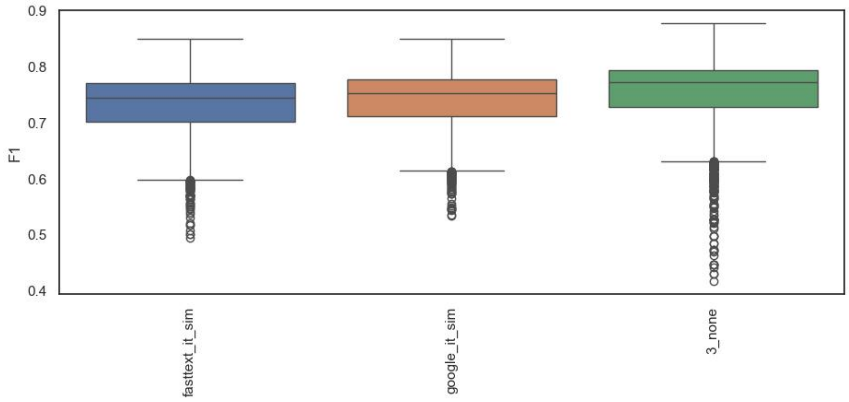


Figure A21: RF on Word Mark Data: Average Performance of Features for Similarity of Goods and Services

RF on Figurative Mark Data

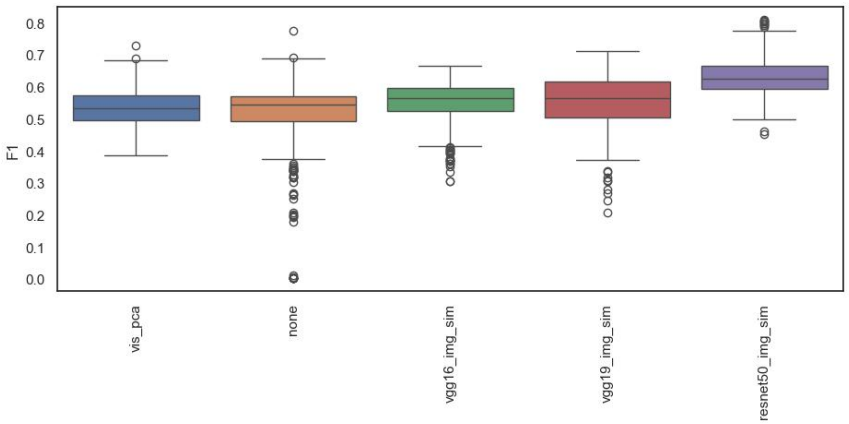


Figure A22: RF on Figurative Mark Data: Average Performance of Features for Visual Similarity

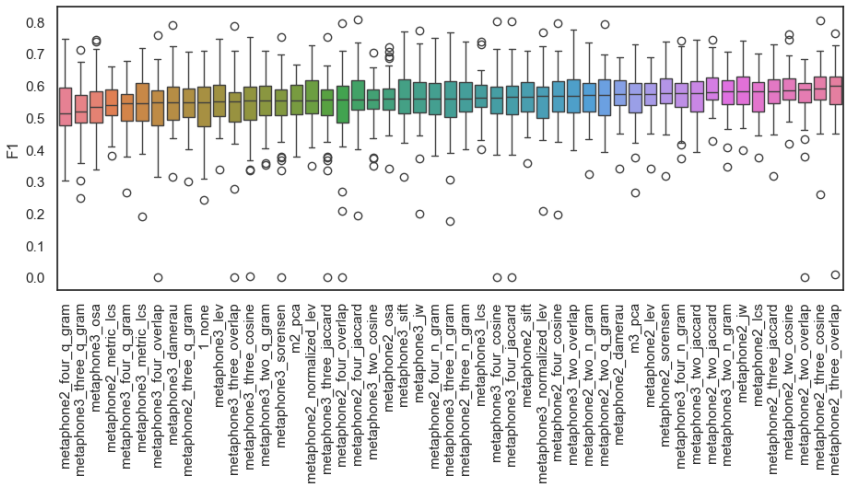


Figure A23: RF on Figurative Mark Data: Average Performance of Features for Aural Similarity

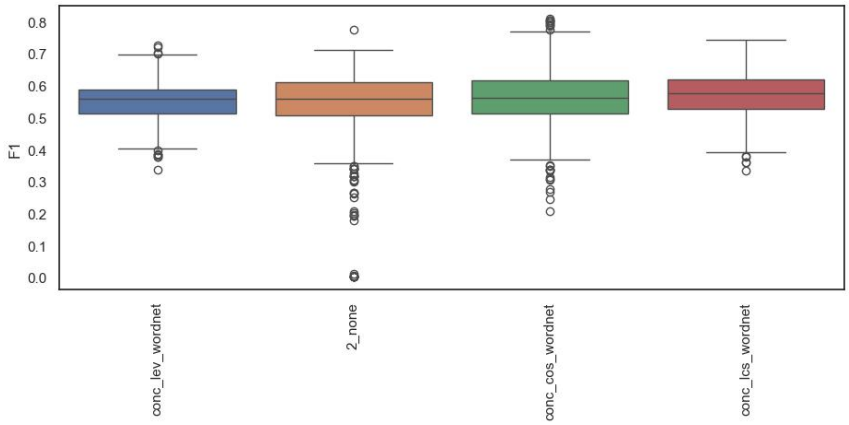


Figure A24: RF on Figurative Mark Data: Average Performance of Features for Conceptual Similarity

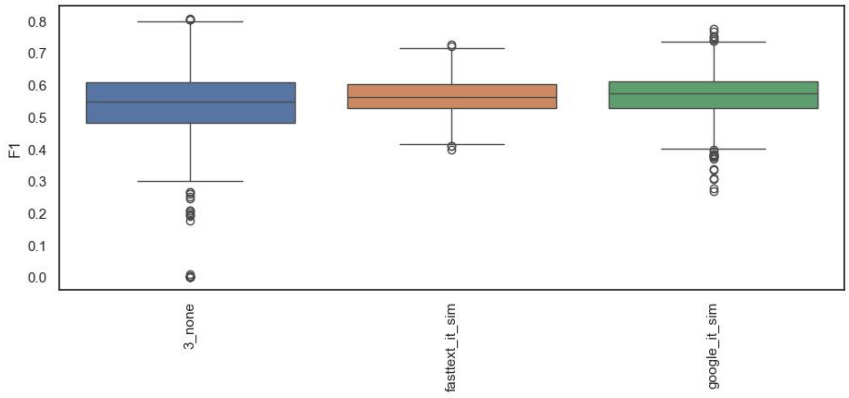


Figure A25: RF on Figurative Mark Data: Average Performance of Features for Similarity of Goods and Services

SVM on Word Mark Data

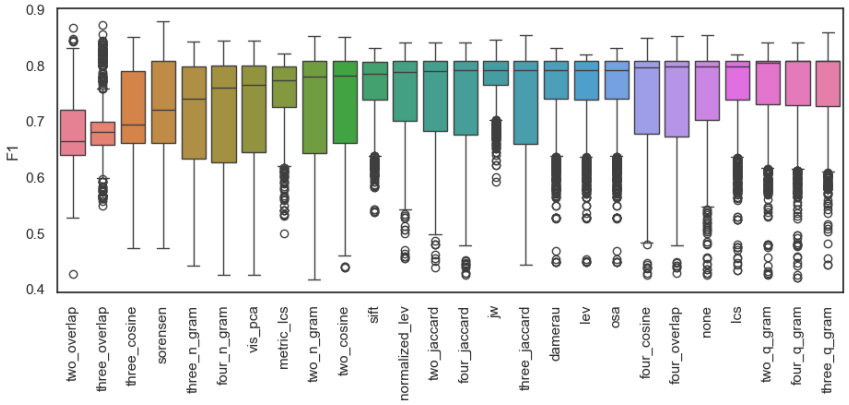


Figure A26: SVM on Word Mark Data: Average Performance of Features for Visual Similarity

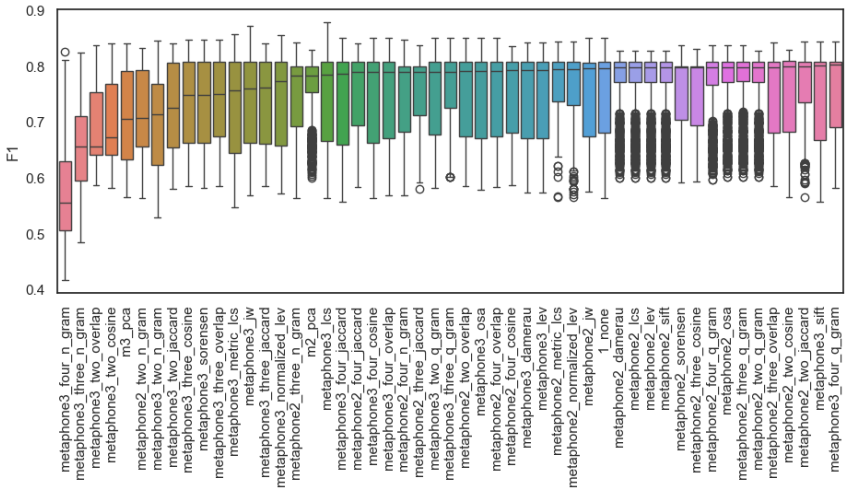


Figure A27: SVM on Word Mark Data: Average Performance of Features for Aural Similarity

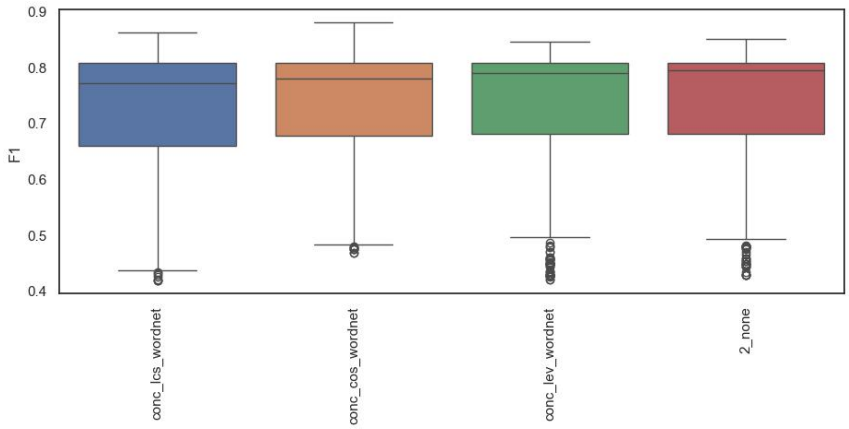


Figure A28: SVM on Word Mark Data: Average Performance of Features for Conceptual Similarity

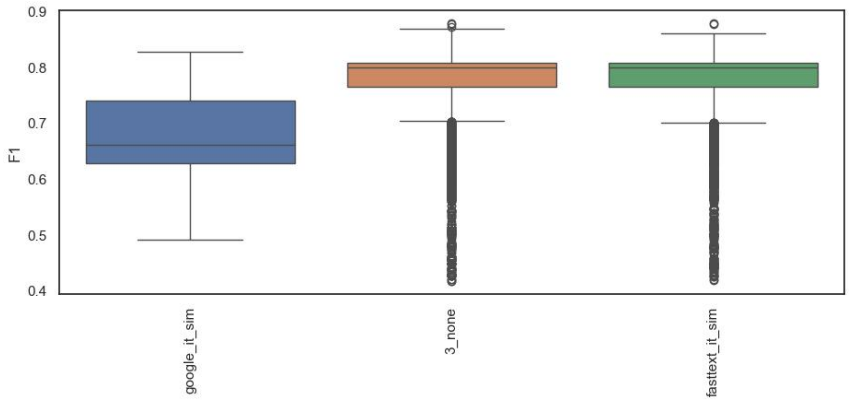


Figure A29: SVM on Word Mark Data: Average Performance of Features for Similarity of Goods and Services

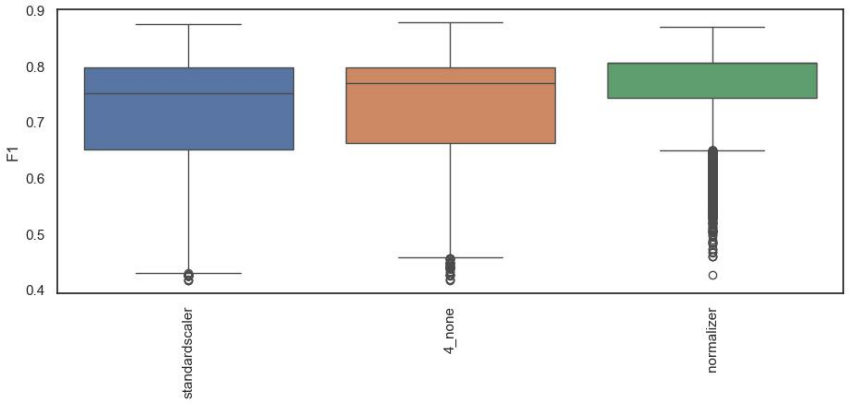


Figure A30: SVM on Word Mark Data: Average Performance of Scalers

SVM on Figurative Mark Data

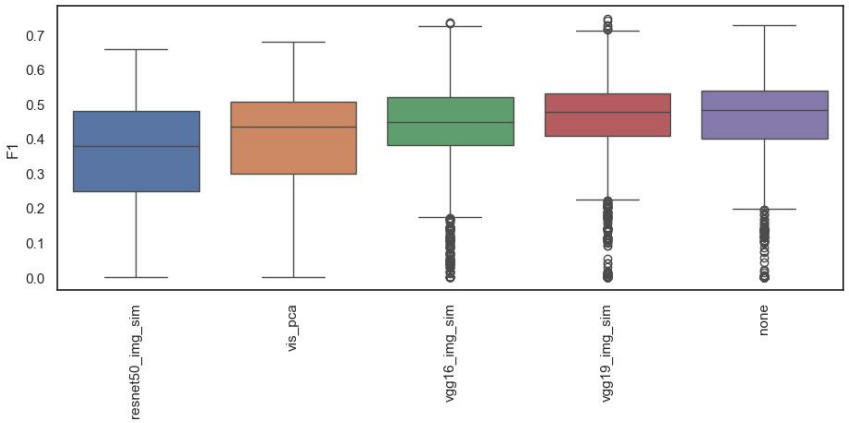


Figure A31: SVM on Figurative Mark Data: Average Performance of Features for Visual Similarity

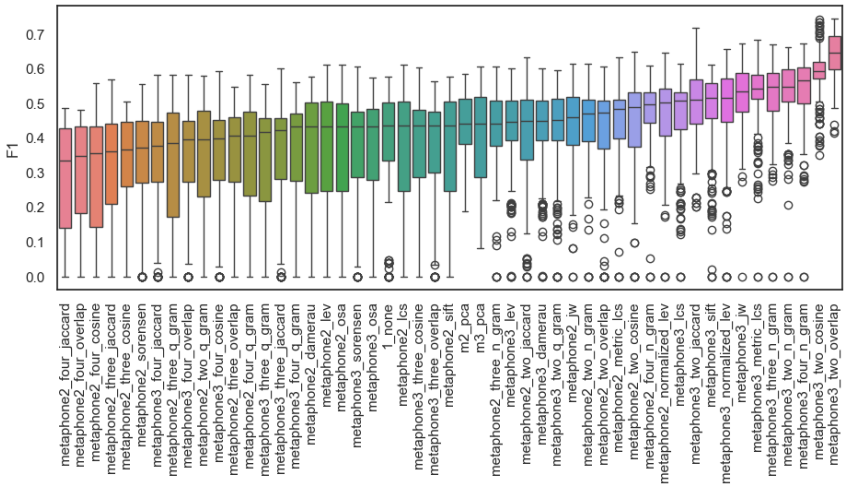


Figure A32: SVM on Figurative Mark Data: Average Performance of Features for Aural Similarity

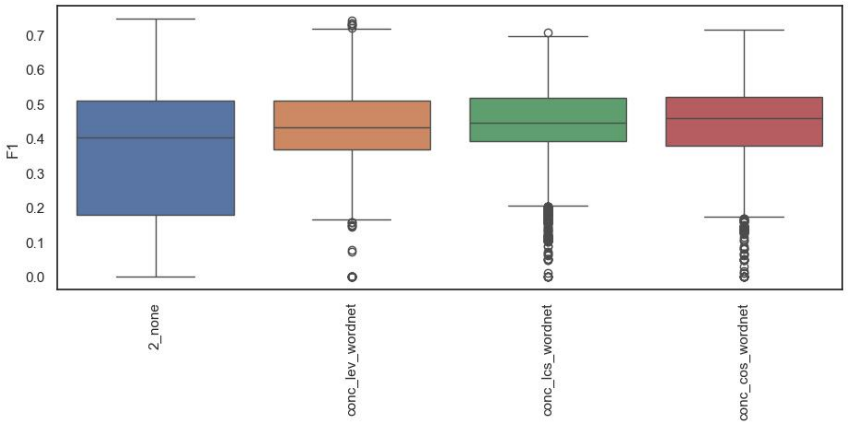


Figure A33: SVM on Figurative Mark Data: Average Performance of Features for Conceptual Similarity

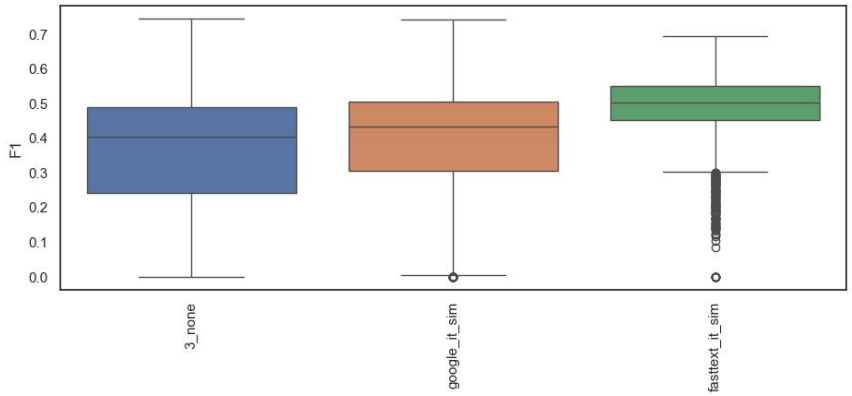


Figure A34: SVM on Figurative Mark Data: Average Performance of Features for Similarity of Goods and Services

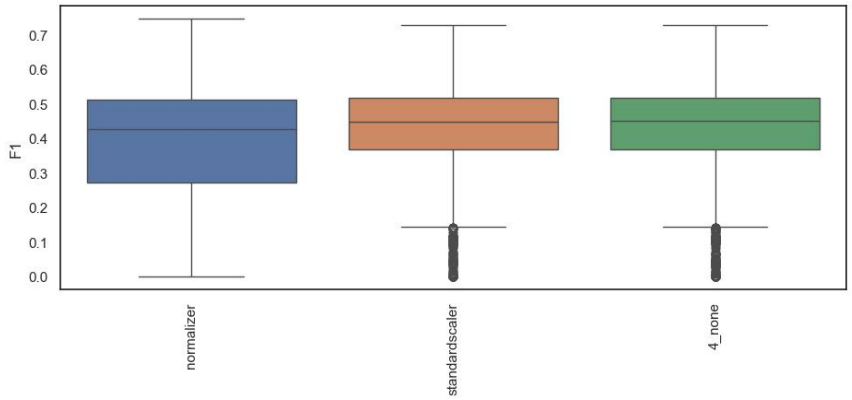


Figure A35: SVM on Figurative Mark Data: Average Performance of Scalers