Automating *Abercrombie*: Machine-Learning Trademark Distinctiveness

Shivam Adarsh, Elliott Ash, Stefan Bechtold, Barton Beebe & Jeanne Fromer[1]

July 21, 2024

Abstract

Trademark law protects marks in order to enable firms to signal their products' qualities to consumers. To qualify for protection, a mark must be able to identify and distinguish goods. U.S. courts typically locate a mark on a "spectrum of distinctiveness" – known as the Abercrombie *spectrum – that categorizes marks as fanciful, arbitrary, or suggestive, and thus as "inherently distinctive," or as descriptive or generic, and thus as not inherently distinctive. This paper explores whether locating trademarks on the* Abercrombie *spectrum can be automated using current natural-language processing techniques. Using about 1.5 million U.S. trademark registrations between 2012 and 2019 as well as 2.2 million related USPTO office actions, the paper presents a machine-learning model that learns semantic features of trademark applications and predicts whether or not a mark is inherently distinctive. Our model can predict trademark actions with 86% accuracy overall, and it can identify subsets of trademark applications where it is highly certain in its predictions of distinctiveness. We further analyze which features in trademark applications drive our model's predictions. We then explore the practical and normative implications of our approach. On a practical level, we outline a decision-support system that could, as a "robot trademark clerk," assist trademark experts in their determination of a trademark's distinctiveness. Such a system could also help trademark experts understand which features of a trademark application contribute the most towards a trademark's distinctiveness. On a theoretical level, we discuss the normative limits of the* Abercrombie *spectrum and propose to move beyond* Abercrombie *for trademarks whose distinctiveness is uncertain. We discuss how machine-learning projects in the law not only inform us about the aspects of the legal system that may be automated in the future, but also force us to tackle normative tradeoffs that may be invisible otherwise.*

---

[1] Student, University of Zurich; Associate Professor of Law, Economics, and Data Science, ETH Zurich; Professor of Intellectual Property, ETH Zurich; John M. Desmarais Professor of Intellectual Property Law, NYU School of Law; Walter J. Derenberg Professor of Intellectual Property Law, NYU School of Law. The authors would like to thank Chris Buccafusco, Christoph Engel, Dev Gangjee, Aniket Kesari, Filippo Lancieri, Mark Lemley, Daryl Lim, Anup Malani, Jonathan Masur, Marc McKenna, Lisa Ouellette, Jason Rantanen, Dominik Stammbach, David Stein, as well as participants at ETH Zurich, the American Law and Economics Association 2023 conference, the Conference for Empirical Legal Studies 2023 and the Intellectual Property Scholars Conference 2024 for helpful feedback. Elias Landes and Simona Ramsperger provided excellent research assistance.

Table of Contents

1. Introduction

In 2022, nearly 800,000 trademark applications were filed with the U.S. Patent & Trademark Office (USPTO) (U.S. Patent & Trademark Office 2022a). For each of these trademark applications, an examining attorney at the USPTO must decide whether the criteria for trademark protection are met. Companies developing a new brand that would like to register their trademarks must predict the outcome of this examination when devising their brand strategy. And courts must evaluate the outcome of the USPTO's examination when deciding trademark registrability or infringement disputes.

Examining a single trademark application can be a time-intensive task. Examining hundreds of thousands of trademark applications can easily clutter an administrative agency. As of May 2023, the average processing time for a new trademark application at the USPTO was 14.6 months, and it took the USPTO about 8.5 months to issue a first office action after a trademark application was filed (U.S. Patent & Trademark Office 2023). Given the costs of these delays for companies and the economy, there is substantial social value in making this process more efficient.

A key component of the trademark application process is determining whether or not a mark is "inherently distinctive," meaning that consumers immediately recognize that the mark, because of its semantic or other qualities, is a designation of the source of the product or service. U.S. trademark law traditionally decides distinctiveness following *Abercrombie & Fitch Co. v. Hunting World, Inc.*, a 1976 decision by the U.S. Court of Appeals for the Second Circuit (Abercrombie & Fitch 1976). The *Abercrombie* spectrum includes five categories. The first three categories indicate inherent distinctiveness: (1) Fanciful marks, which are completely invented terms like "Xerox," have the highest level of distinctiveness; (2) Arbitrary marks, such as "Apple" for computers, have no direct link to the product; and (3) Suggestive marks, like "Ivory" for soap, imply something about the product. The latter two categories are not inherently distinctive: (4) Descriptive marks, for instance "Pizza Hut" or "iPhone," directly describe the product; (5) Generic marks, like "Escalator" or "Aspirin," are common names for the products themselves. The *Abercrombie* determination is important in trademark applications because if a mark is deemed not to be inherently distinctive, the mark will register only if the applicant additionally shows that the mark has acquired distinctiveness, typically through advertising and use of the mark in the marketplace. This showing is often difficult and costly to make.

This paper explores the extent to which natural language processing techniques can help to automate *Abercrombie* determinations. We present a machine-learning pipeline that is trained on the text of trademark applications to predict whether a trademark is inherently distinctive. As the training dataset, we use the one million trademark applications filed with the USPTO between 2012 and 2017. We then validate and evaluate the out-of-sample performance using the 234,000 trademark applications filed in 2018 and the 264,000 trademark applications filed in 2019. The

model takes as input the mark and associated application data, and outputs a predicted probability that the mark is inherently distinctive for the associated product or service.

We adapt modern deep-learning-based machine learning methods for text (Devlin et al. 2019). Our method is context-sensitive, in the sense that it works not just by counting individual words but by learning how words are connected, thereby aiming at a semantic understanding of trademark applications. In our approach, the model input includes the mark, the product description, and the product class. From the mark text, the model can directly learn fancifulness – i.e., made-up words such as "Xerox". From the contextual connections between the mark and the product description and Nice class, the model can learn the more subtle category of arbitrariness – that is, mark terms that are in the dictionary yet semantically orthogonal to the product attributes (e.g., "Apple computer").

We find that, in the aggregate, our model's binary predictions agree with the trademark office's decisions about 86% of the time (AUC = 0.77). The relatively high error rate of 14% reflects, in part, that *Abercrombie* is a subjective determination where humans often disagree, so a significant error rate is unavoidable. However, we also find that the model produces a reliable confidence score for distinctiveness – that is, a predicted probability for various USPTO outcomes. For trademarks where the model has high confidence – i.e., the predicted probability is close to 0% or close to 100% – the precision of the model decision is very high (above 90%). We provide a number of robustness checks to show that our approach is fit for purpose and preferred to alternatives.

We argue that the varying confidence in predicting *Abercrombie* is an important feature of our model for trademark theory and practice. On a practical level, we propose a "robot trademark clerk," which would provide a decision-support system for trademark examiners, but also judges, attorneys, and firms who are interested in whether a particular mark is inherently distinctive. Such a system could also help trademark experts understand which features of a trademark application contribute the most towards a trademark's distinctiveness. On a normative level, we discuss whether the varying confidence in predicting *Abercrombie* should lead us to look for new normative foundations on whether to grant protection to middle-ground trademarks, whose inherent distinctiveness is unclear. These points suggest that applying machine-learning methods to trademark law may not only inform us about the extent to which trademark procedure can be automated. It may also shed light on normative tradeoffs that trademark theory must engage in that were not visible without the assistance of machine learning.

The paper proceeds as follows. Section 2 provides the relevant legal background on U.S. trademark law and registration procedure. It also gives an overview of the existing literature that applies machine-learning methods to trademark law or thinks about such applications in theoretical terms. Section 3 presents our approach towards predicting whether a mark is

inherently distinctive on the *Abercrombie* spectrum, the data we use for our approach, and the concrete implementation of our approach in a natural-language processing setting. Section 4 presents the results of our model predictions and explores which factors in the trademark application drive our model's predictions. Section 5 discusses how our framework could be used to develop a robot trademark clerk, and how it sheds light on normative limitations of the Abercrombie spectrum. Section 6 discusses limitations of our framework and next steps to overcome them. Section 7 concludes the paper.

2. Legal Background and Related Literature

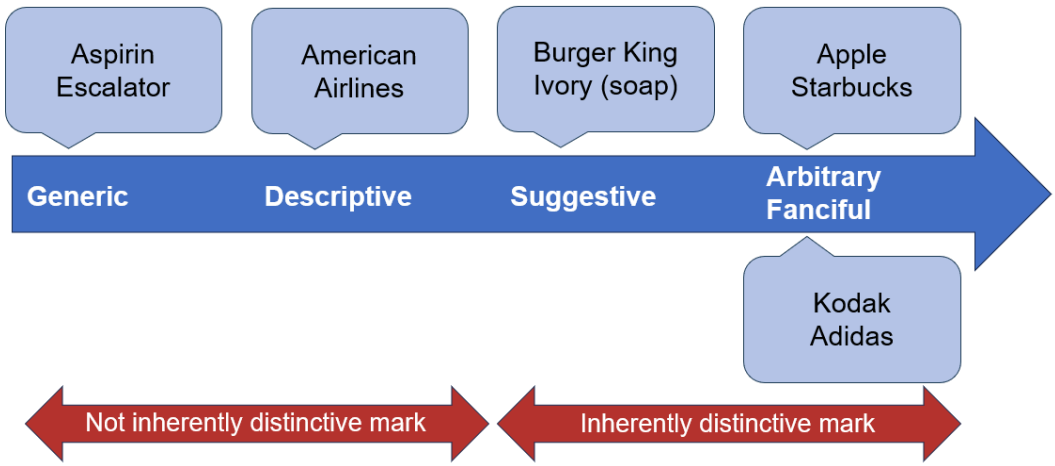
2.1 Legal Background

When a trademark registration application is filed with the USPTO, an examining attorney reviews it to determine whether it meets all the criteria for registration. Among other criteria, the examining attorney determines whether the mark is descriptive or deceptive, and whether conflicting marks are already registered. If the trademark application does not fulfill all registration criteria, the USPTO issues an office action to the applicant, listing the reasons for refusal. The applicant can then respond, and the USPTO will then either proceed with the application or issue a final office action refusing registration. If all the registration criteria are met, the examining attorney approves the application for publication in the official journal of the USPTO ("Official Gazette"). Competitors or other parties who believe they could be harmed by the mark's registration can then file an opposition against it, which will be decided by the Trademark Trial and Appeal Board. If there is no opposition or the opposition fails, the USPTO will register the trademark (Gilson 2023: §16.01; U.S. Patent & Trademark Office 2022b: §704.1). Once a trademark is registered in the USPTO's Principal Register, the registration is *prima facie* evidence of the mark's validity, including its distinctiveness (Gilson 2023: §2.05[2]).

As part of the substantive examination, the examining officer must determine whether a mark is able to "identify and distinguish [...] goods [...] from those manufactured or sold by others and to indicate the source of the goods" (15 U.S.C. § 1127). Conceptually, trademark law follows a classification developed by Judge Friendly in the *Abercrombie* decision (Abercrombie & Fitch 1976). The *Abercrombie* spectrum classifies trademarks according to their degree of inherent distinctiveness: (1) fanciful marks, which have the highest degree of inherent distinctiveness, are coined terms (e.g., Xerox for copiers); (2) arbitrary marks, which have no semantic connection to the relevant product (e.g., Apple for computers); (3) suggestive marks, which suggest or are metaphorically related to product characteristics (e.g., Ivory for soap); (4) descriptive marks, which describe product characteristics (e.g., Coca-Cola a cola drink or iPhone for a mobile phone); and (5) generic marks, which refer to the type of product (e.g., Escalator or Aspirin, Beebe 2006: 1634; Beebe & Fromer 2018: 957; see Figure 1). Arbitrary and fanciful marks are understood to be the most highly inherently distinctive, followed in decreasing order by suggestive marks, descriptive marks, and generic marks. Courts rely on a variety of evidence to

determine whether a mark is distinctive, including consumer surveys, dictionary definitions, expert testimony, and evidence of use over time (Gilson 2023: §2.05[1]).

Figure 1: *Abercrombie* Spectrum: From Generic to Fanciful Marks



Trademark law treats marks that are fanciful, arbitrary, or suggestive as inherently distinctive marks, as consumers are understood to immediately interpret them as designations of source. By virtue of their inherent distinctiveness, such marks can be registered on the Principal Register – provided they meet all other requirements for registration. For descriptive marks, trademark law treats them as not inherently distinctive, as consumers may interpret them as mere product descriptions. Therefore, the trademark applicant must produce evidence that the mark is distinctive of source to consumers (so-called "secondary meaning" or "acquired distinctiveness"). A descriptive mark may acquire distinctiveness over time through use, in particular if consumers learn to recognize the mark as an indication of source (Beebe & Fromer 2018: 957-58). As a general matter, generic marks cannot be registered as trademarks at the USPTO (on the distinction between descriptive and generic marks after the Supreme Court's *Booking.com* decision, see Fromer 2022).

If a mark does not achieve distinctiveness – either inherently (for fanciful, arbitrary, or suggestive marks) or through "secondary meaning" (for descriptive marks) – it cannot be registered or otherwise protected as a trademark. While inherently distinctive trademarks that register are placed on the Principal Register (generating *prima facie* evidence of the mark's validity, as mentioned before), descriptive trademarks with no secondary meaning can only be registered on the Supplemental Register (providing only limited notice and standing functions; 15 U.S.C. §1091; U.S. Patent and Trademark Office 2022b: §1209.1; Gilson 2023: §3.05[3], §4.07[5])). Once a mark published only in the Supplemental Register has acquired secondary meaning, it can be published in the Principal Register.

Determining the distinctiveness of a mark is not only important for determining whether a mark is protectable. It is also important because the scope of trademark protection is proportional to the inherent distinctiveness of a mark, which can be an important factor in determining the likelihood of confusion in trademark infringement litigation, along with factors such as the extent of marketplace use and advertising of the mark (Beebe 2006: 1634, 1637; Gilson 2023: §2.01[6]). For the purposes of trademark registration, however, the five-tier *Abercrombie* spectrum boils down to a simple dichotomy: those marks which are distinctive of source and those marks which are not. Among the former group of marks, some marks are inherently distinctive (fanciful, arbitrary and suggestive), while others may acquire distinctiveness through secondary meaning (descriptive; Beebe 2004: 671; Gilson 2023: §2.01[6]).

2.2 Literature

This paper builds on various strands of literature. Tushnet (2017: 870-71) points out that the procedural and substantial aspects of the trademark registration process are typically neglected in trademark law scholarship. With our study, we hope to narrow this research gap. Various scholars have explored the relationship between trademark registrations and language. Beebe (2006) analyzes federal district court opinions from 2000 to 2004 to determine which multifactor tests courts use when determining likelihood of consumer confusion. Beebe & Fromer (2018) measure the extent to which the most common English words and syllables are registered as trademarks, thereby potentially depleting the supply of possible marks. On a conceptual level, Hemel & Ouellette (2021) argue that such depletion can create two distinct types of costs for the trademark system: proximity costs (where different firms use similar marks) and distance costs (where firms eschew these proximity costs by using marks that are difficult for consumers to remember).

Scholars, trademark offices and the consulting industry have started to explore the implications natural-language processing tools have for the trademark system (for related attempts in the patent system, see Hain et al. 2022; Ashtor 2022). Often, legal scholars resort to debating the impact of artificial intelligence on the trademark system on a conceptual level (Moerland & Freitas 2021; Katyal & Kesari 2020; Gangjee 2021; Lim 2022). Trademark offices have started to explore systems classifying trademark applications into the respective Nice classes of goods and services (such as apparel, computer software, and beverages), to identify similar marks and to determine whether a sign is descriptive (Gangjee 2021: 178-184; Moerland & Freitas 2021: 226). The consulting industry has started to offer related trademark-search and decision support systems (see Katyal & Kesari 2020). Detailed information about the inner workings of such systems are often scarce, however.

Data scientists have applied machine-learning tools to develop a recommendation system identifying semantically similar trademark litigation judgments for a trademark of interest (Trappey et al. 2020). Shackell & De Vine (2022) use word-embedding metrics to analyze whether trademarks have become generic. Showkatramani et al. (2019) use natural language

processing methods to automate manual efforts in classifying trademarks into the 45 classes of the Nice classification system of goods and services.

Most relevant to our paper are projects that attempt to determine trademark distinctiveness with existing data. Even before the big data revolution, Ouellette (2014) proposed to determine trademark strength through conducting Google searches. She argued and demonstrated through case studies that the stronger a trademark is, the more top search results for the trademark will appear on Google. More recently, Goodhue & Wei (2023) explored whether a large-language model such as GPT 3.5 could be used to classify trademarks along the *Abercrombie* spectrum. In a case study of 24 trademarks with some prompt engineering and limited fine-tuning, they report that GPT 3.5 did a mediocre job in distinguishing descriptive from suggestive, fanciful and arbitrary marks. To determine whether their 24 benchmark trademarks are at least suggestive or merely descriptive, they use the fact of whether the trademarks are registered on the Principal Register or the Supplemental Register as a proxy. In their best specification, they find that GPT 3.5 determines 6 out of 15 on the Principal Register as at least suggestive and and 7 of 9 marks on the Supplemental Register as descriptive. In a follow-on study, Goodhue & Xing (2023) use GPT to identify special issues in trademark applications, such as personal names or geographically descriptive marks. They then use a logistic regression model applied to TF-IDF features to classify trademarks on the *Abercrombie* spectrum. Guha et al. (2022) take a broader approach and build a legal benchmark suite for language foundation models – such as GPT – that would enable comparisons on how various foundation models perform on these tasks. One of the over 170 available tasks is to locate a trademark on the *Abercrombie* spectrum (LegalBench 2023). Their initial results provide an F1 score of .42 for the best-performing foundation model (GPT-3 davinci, Guha et al. 2022: 8; see also Guha et al. 2023).

Overall, while data scientists and legal scholars have begun to apply machine-learning methods to select issues in trademark law, to the best of our knowledge no scientific work has developed natural-language processing tools to determine trademark distinctiveness on the *Abercrombie* spectrum on a large scale.[2] More specifically, compared to the existing empirical literature on the *Abercrombie* spectrum, this paper is the first to train a machine-learning model on actual trademark decisions at a large scale, using a dataset with over 1.5 million trademark registrations, while earlier contributions are typically based on querying a standard language model or training a language model with about 3,000 trademark registrations only. In addition, by combining trademark registrations with information on USPTO office actions, we present a clean dataset that enables us to accurately determine whether the USPTO has classified a trademark as distinctive or not (for details, see Section 3.2). Finally, our research design enables us to provide information on how confident our model's predictions are, and we can show which components of trademark applications are driving factors in the model predictions (for details, see Sections

[2] The trademark offices of Australia, the European Union and Singapore have reportedly worked on systems to assist distinctiveness determination, see Gangjee (2021): 184. However, information on these systems is scant. Also, the *Abercrombie* test, which is the focus of this study, does not apply in these jurisdictions.

4.2 and 4.3). All of these features, which we add to the existing body of literature, have important implications for the trademark discourse (see Section 5).

3. Methods and Data

3.1 Predicting *Abercrombie*

In this paper, we want to use natural-language processing techniques to predict whether a USPTO examining attorney would determine that a particular mark is inherently distinctive or not. We want to use existing trademark registrations (including, e.g., the name of the trademark, its Nice class, and its product description) to train a classifier to predict distinctiveness. As we are interested in exploring the semantic features of registered trademarks, we treat only inherently distinctive marks as "distinctive" and do not include in this category descriptive marks that may acquire or have already acquired secondary meaning among consumers (on this distinction, see supra Section 2.1). While the *Abercrombie* spectrum can be used to determine the strength of a trademark on a five-tier scale, for the purposes of trademark registration, our task gets reduced to a simple dichotomy (see supra Section 2.1): distinguishing inherently distinctive marks (either fanciful, arbitrary, or suggestive) from non-inherently-distinctive marks (either descriptive or generic, regardless of acquired secondary meaning).

In order to train our classifier whether or not a mark is inherently distinctive, we ideally would need a comprehensive dataset of all active trademarks with information about their validity, location on the *Abercrombie* spectrum, product description, and Nice class. Two types of potential datasets come to mind. First, one could explore litigated trademark cases. While trademark scholars have coded and analyzed trademark court decisions empirically (see Beebe 2006), such datasets are unfortunately not well-suited for our purposes. The most important limitations are that courts very often do not reveal in their decision what the trademark's registration number is, in which Nice classes the trademark is registered, and where exactly on the *Abercrombie* spectrum the trademark is located (Beebe 2006: 1635-1636). Even if one matched litigated trademarks to trademark registration data, where such information is available, this would not be sufficient. The number of court decisions dealing with trademark distinctiveness is limited and would not suffice to train machine-learning models which typically require very large training datasets.[3]

A second potential dataset is trademark registration data. The advantage of trademark registration data is that all of the information about trademarks mentioned above is readily available from the USPTO (Graham et al. 2013). When a USPTO examining attorney decides to register a trademark, the registration record does not directly reveal whether the trademark is inherently distinctive (as also descriptive trademarks with acquired secondary meaning can be registered). Rather, one needs to use a proxy to determine the inherent distinctiveness of a trademark. In their case study, Goodhue & Wei (2023) employ as a proxy whether a trademark

[3] In Appendix A.2, we present and analyze such a hand-coded dataset for validation purposes.

has been registered on the Principal Register or the Supplemental Register. Mapping inherently distinctive trademarks to the Principal Register and descriptive trademarks to the Supplemental Register is imprecise, however. As mentioned in Section 2.1, a descriptive trademark that has acquired secondary meaning can and will be registered on the Principal Register. Therefore, in our view, using the Principal vs. Supplemental Register as a proxy to determine the inherent distinctiveness of a mark is problematic.

In this paper, we develop a distinctiveness indicator using a combination of publications of trademarks in the Primary and the Supplemental Register, USPTO office actions, and data available through USPTO case files. We start by considering the publication[4] of a trademark application in the USPTO Official Gazette as a proxy for inherent distinctiveness. As explained in Section 2.1, when a trademark application is filed with the USPTO, an examining attorney conducts a complete examination, including whether or not a trademark is descriptive. If the examining attorney decides that the trademark is descriptive and lacks secondary meaning, the application will be refused through an office action. If, however, the examining attorney decides that all the registration criteria are met, the trademark application is published in the USPTO Official Gazette. Thereafter, a competitor can oppose the mark's registration if, for example, the competitor holds rights in a confusingly similar mark. If there is no opposition or the opposition fails, the USPTO will register the trademark. The trademark's publication provides information about the examining attorney's assessment of the trademark registrability, including the trademark's distinctiveness: if the examining attorney finds a mark merely descriptive and lacking in acquired distinctiveness, the application will not be published in the Official Gazette.

Unfortunately, whether a trademark gets published in the Official Gazette is an imperfect proxy for whether the USPTO examining officer treats a mark as inherently distinctive (publication: yes; no publication: no). First, as mentioned in Section 2.1, descriptive marks may also be published in the Official Gazette if the applicant has provided evidence of secondary meaning. Second, marks that are inherently distinctive may not get published in the Official Gazette for other reasons (e.g., because they are confusingly similar to an already-registered mark). Therefore, if we took publication in the Official Gazette as an indicator of inherent distinctiveness, our analysis might suffer from substantial false positive and false negative errors.

To remedy such errors, we treat trademarks that get published for registration on the Principal Register as inherently distinctive, but exclude trademark applications for which the applicant did not claim inherent distinctiveness, but provided evidence of secondary meaning (so-called "2(f) applications," see 15 U.S.C. §1052(f)). We also treat trademarks as inherently distinctive if they did not get published but, according to their corresponding office actions, did not get refused due

[4] By focusing on trademark publication rather than trademark registration, we do not have to worry about oppositions from competitors with regard to confusingly similar trademarks. As we are interested in how USPTO examining attorneys determine trademark distinctiveness, focusing on the first step in the trademark registration process – publication in the USPTO Official Gazette – provides a cleaner distinctiveness indicator.

to a lack of distinctiveness (for example, they are inherently distinctive, but get rejected on other grounds, such as 15 U.S.C. § 1502 (d), for being confusingly similar with an already-registered mark).

By contrast, we treat trademarks published for registration on the Principal Register that resulted from a "2(f) application" as not inherently distinctive marks. Trademark applications that do not get published and, according to their corresponding office actions, get refused due to a lack of distinctiveness (even if they also get refused on other grounds as well) are treated as not inherently distinctive marks as well. Finally, trademarks published in the Supplemental Register are also not inherently distinctive.

In our view, combining information about which trademarks get published in the Principal and the Supplemental Register with information from USPTO office actions and case files is the most accurate way of creating a training data set for the distinctiveness threshold on the *Abercrombie* spectrum. Table 1 provides an overview of our categorization. We use this categorization, coupled with all of the other features that can be retrieved from the application about the trademark, to train a classifier. If we provide the system with a trademark with certain features, the system should then be able to predict whether or not the trademark is inherently distinctive.

Table 1 : Creation of Distinctiveness Indicator

Trademark Type	Explanation	How to Detect in the USPTO Data	Outcome Variables in the USPTO Data
Inherently distinctive	all trademarks that get published in the Principal Register WITHOUT 2(f) trademark applications (not inherently distinctive, but with secondary meaning)	Published in the Principal Register BUT FOR 2(f) applications	Publication Indicator = True AND Acquired Distinctiveness Indicator = False
	all trademark applications that do not get published and, according to their corresponding office actions, do not get refused due to a lack of distinctiveness (they are inherently distinctive, but get rejected on other grounds, e.g. 2(d))	Not published in the Principal or Supplemental Register BUT FOR rejections due to mark being descriptive (2(e)(1)) or generic (according to office action data)	Publication Indicator = False, Supplemental Register Indicator = False, AND (from office actions), Descriptive Sum = 0 AND Generic Sum = 0
Not inherently distinctive	all trademarks that get published in the Principal Register which came from a 2(f) trademark application (not inherently distinctive, but with secondary meaning)	Published in the Principal Register AND 2(f) application status	Publication Indicator = True, AND Acquired Distinctiveness Indicator = True
	all trademark applications that do not get published and, according to their corresponding office actions, do get refused due to a lack of distinctiveness (even if they also get refused on other grounds as well)	Not published in the Principal AND rejections due to mark being descriptive (2(e)(1)) or generic (according to office action data)	Publication Indicator = False, AND (from office actions), Descriptive Sum is non-zero OR Generic Sum is non-zero
	all trademark applications that do get published in the Supplemental Register (because they lack inherent distinctiveness).	Published in the Supplemental Register	Supplemental Registration Indicator = True

3.2 Data

As outlined in the preceding section, we use the USPTO Trademark Case Files Dataset (Graham et al. 2013) as our central data source. This dataset includes records on 12.4 million trademark applications and registrations associated with the USPTO. The earliest entries in the dataset go back to 1870. The dataset originates from the primary USPTO database responsible for managing trademarks. It encompasses a wide range of information, including details about the characteristics of marks, prosecution events, ownership, classification, third-party oppositions,

and the history of renewals. As a result, the dataset offers an extensive labeled data source for analyzing trademark publication, registration and acquired distinctiveness.

We also use a dataset developed by Beebe and Fromer (Beebe & Fromer 2020, used in Beebe & Fromer 2018) of all trademark office actions that the USPTO issued from 2003, when it began posting office actions online, through 2019. For applications filed from 2012 through 2019, the dataset includes the full text of 2.2 million office actions issued by USPTO trademark examining attorneys concerning 1.5 million trademark applications. We autocoded each office action according to certain keywords and phrases that trademark examining attorneys consistently use when refusing registration based on descriptiveness or genericism.

To streamline the analysis and prioritize recent data, we use data from 2012 through 2019 in our study. To simplify the analysis and focus on text features of trademarks, we drop marks with images (about 20% of the sample). For similar reasons, we also drop marks with five or more words (about 6% of the sample, Appendix Figure B1 shows a histogram of word lengths).

After preprocessing, we divide the data into training, validation, and test sets. For training, we use trademarks that have been filed between 2012 and 2017. We use trademarks filed in 2018 as our validation dataset, and trademarks filed in 2019 as our test dataset. Overall, we have about one million data points for training, about 234,000 data points for validation and about 264,000 data points for testing purposes.

Our primary outcome to be predicted is an indicator variable for "distinctiveness," indicating whether the trademark application had a distinctive mark or not, as created in Table 1. We assign 1 to trademark entries which are inherently distinctive, and 0 otherwise. Approximately 84% of all trademark applications which we use as our test dataset are classified as inherently distinctive trademarks. The incidence rate for all outcome labels are provided in Appendix Table B1.[5]

For each trademark, we use the following variables in our predictions (where features 1 to 4 come directly from the USPTO Trademark Case Files Dataset):

1. Statement Text: This provides a description of the proposed mark phrase and the product to which it is supposed to be affixed. As recently developed language models can use all the information from input text, including punctuations and stopwords, we do not preprocess the statement text.
2. Mark: This lists the proposed mark.

[5] For robustness, we have a set of alternative outcome labels that summarize other aspects of the USPTO process. First, if a trademark entry has a registration date present, the registration indicator is 1, else 0. The incidence rate for registration is 58.1%. Second, the acquired distinctiveness indicator is already present in the case files dataset provided by USPTO, but it is a rare outcome at 1.8%. Third, we create a variant of the registration indicator where we filter out Intent to Use applications (variant 1 for registration indicator). Fourth, we predict registration but dropping rows where a mark has acquired distinctiveness (variant 2 for registration indicator). Fifth, we produce a variant of acquired distinctiveness but dropping rows where a mark has been registered (variant 1 for acquired distinctiveness). The results for these variants are shown in Appendix A.4.

3. Nice classes: This lists the class of goods and services (1-45) for which the trademark is applied.
4. Pseudo mark: If a trademark has an alternative meaning or spelling, the Trademark Case Files Dataset provides such spelling. “Pseudo mark” entries, which are not displayed in search results and are not part of the official application or registration, are added by the USPTO in cases where spellings exist that are very similar or phonetically equivalent to the mark. If, e.g., a trademark includes “4U”, the term “for you” would be added to the pseudo-mark field. Or, LIFT could appear as a pseudo-mark entry to the trademark LYFT (see U.S. Patent & Trademark Office 2022b: §104).
5. Mark length: This counts the number of words in the trademark.
6. Translated mark: In case of trademarks in a foreign language, we translated the trademark to English, using Google’s Translate API. Translations from different languages can give some sense if a mark can be published or not. Our assumption relies on the fact that applicants can use a language other than English to publish their marks, and create a brand name. Translating these marks to English gives our model an understanding of the underlying meaning of the mark.
7. Dictionary indicator: We use a binary indicator to signify whether any word from the trademark is present in the WordNet Dictionary (Princeton University 2023). This provides some indication whether a trademark is a standard English or a made-up word.
8. Nice class description: This describes the relevant Nice class of the trademark in a broad way (European Union Intellectual Property Office 2023). We focus on high-level Nice classes only (1-45).

Since there are different statement type codes for each trademark (each trademark is represented by a serial number), we aggregate the data on a trademark level by concatenating different statement texts available for a given trademark. We do this for all different statement types, except for pseudo marks. As pseudo marks can be thought of as “wordplay” on trademarks, we keep them as a separate feature. By including pseudo marks, we provide our model with some information on how humans understand a particular word from a linguistic perspective, thereby enriching our model’s understanding of context. A summary of the features used is presented in Table 2.

Table 2: Summary Statistics of Features

Feature	Feature Type	Additional Information	
Statement text	Text	Average number of words: 28.87	
Mark	Text.	Average number of characters: 12.60	
Nice classes	Categorical	Nice class distribution is shown in Figure B2	
Pseudo mark	Text	Pseudo mark present for 27.84% marks	
Mark length	Numerical	Number of words in mark	% datapoints
		1	41.08%
		2	31.92%
		3	18.40%
		4	7.60%
Translated mark	Text	14.42 % marks translated	
Dictionary indicator	Binary Indicator	Present for 58% marks	
Nice class description	Text	Average number of words: 22.42	

3.3 Text Classification

As described in Section 3.1, we want to be able to provide a machine-learning model with a trademark with certain features so that the model predicts whether or not the trademark is inherently distinctive (according to our indicator presented in Table 1). In choosing our machine classification method, we are guided by the importance trademark practice puts on context. In particular, we are looking for methods that enable the classifier to get a semantic understanding of trademark applications.

We are interested in using a context-sensitive neural net that learns semantic features of trademark applications, along with text representations of covariates. We follow a text classification approach using a BERT-based model. BERT (Devlin et al. 2018) is a large pre-trained transformer model that learns to understand language by guessing left-out words in short documents. Through this process, BERT obtains a semantic understanding of language, where words are understood in the context of other surrounding words. That is, BERT is preferred to other approaches because it not only counts words and phrases, but learns first-order and second-order connections between words, even at a distance.

BERT models can then be applied to many downstream tasks such as classification, question-answering, or summarization. In our case, BERT can learn subtle connections between the different words in the text input, such as how the words in a mark are related to the words in the product description or class description.

BERT takes only text as inputs. We concatenate all the features of a trademark application to form a paragraph, which acts as an input to the BERT model. Figure 2 shows the input format using an example ('ZIPSCENE').[6]

Figure 2: Illustration of Bert Text Classifier Input

[CLS] {Mark} [SEP] {Mark Description from USPTO (Preprocessed and Concatenated)} [SEP] {English translation is translation/ no translation required} [SEP] mark {present/absent} in Wordnet [SEP] mark length is {number of words} [SEP] NICE category is {nice category} [SEP] {Nice class description} [SEP] Pseudo mark is {pseudo mark string/absent}

(a) Input Format for BERT

[CLS] ZIPSCENE {Mark} [SEP] Marketing and market research services; Analyzing and compiling business and marketing data, namely, data assembly, integrations, and analysis relating to customer behavior and transactions; Customer relationship consultation, namely, developing strategies and programs for communicating with customers {Mark Description from USPTO (Concatenated)} [SEP] no translation required [SEP] mark absent in Wordnet [SEP] mark length is 1 [SEP] NICE category is 35 [SEP] Advertising; Business management, organization and administration; Office functions {Nice class 35 description} [SEP] Pseudo mark is ZIP SCENE

Label: Distinct

(b) Example: ZIPSCENE

Note. Panel (a) shows the input format for BERT. Panel (b) illustrates an example of ZIPSCENE following the format described in (a) along with its label. [CLS] and [SEP] are special tokens used in BERT to signify start-of-sequence and sentence separation respectively. For RoBERTa, we replace the [CLS] token <s> and the [SEP] token by </s>.

We match these text inputs with the associated label to form a dataset. We then take a pre-trained BERT model and feed in the new pairs to fine-tune it on the new classification task: whether or not the associated trademark is distinctive (according to our indicator presented in Table 1).[7]

Choosing a threshold for a classification objective is subjective, and depends on the incidence rate of the dependent variable, as well as on the fact that output probabilities truly estimate the data distribution. To support a decision-making framework using our fine-tuned BERT model, it

[6] We also compare our main results with Xgboost (Chen & Guestrin 2016), a more classical machine-learning model. Our fine-tuned DistilBERT outperforms XgBoost model for all the standard evaluation metrics used for classification tasks. Additional details for Xgboost can be found in Appendix A.7.

[7] We trained the model for two epochs with a batch size of 16. No layers were frozen during the training, which allowed the fine-tuned model to learn the word representations in a slightly different way than the original DistilBERT.

is important that the output probabilities are reliable, and correctly estimate the true correctness likelihood. Hence, we calibrate the output probabilities to reflect an estimate of the data distribution (Guo et. al 2017). The resulting adjusted model will give predicted probabilities that reflect the actual outcome rates in the data.

In our study, the specific model variant used is DistilBERT, an efficient and fast BERT implementation (Sanh et al, 2020). As a comparison to DistilBERT, we report results for two other models. As a more capable model, we use RoBERTa (Liu et al. 2020), a larger BERT-style model that has been pre-trained on more data than BERT. As a classical-machine-learning baseline (that is, not using deep learning), we implement XGBoost, which is an ensemble of decision trees that vote on the outcome based on the text inputs. This baseline is explained in detail in Appendix A.7.

4. Results

This section reports the results from predicting whether a trademark is inherently distinctive.

4.1 Model Performance

Table 3 presents the results obtained from our models on the test dataset. We report the following classical machine-learning evaluation metrics: Accuracy means the proportion of correct predictions. Balanced accuracy means the average recall (accuracy) for the two classes. Precision is the accuracy conditional on prediction; F1 is the geometric mean of precision and recall. Finally, AUC (Area Under the ROC Curve) gives the accuracy with which the model can rank two randomly selected observations by their predicted probabilities across classes. The first column provides the weakest baseline of guessing the modal class (distinctive). Accuracy is the proportion of the modal class, and balanced accuracy and AUC are both 0.5. This shows the minimum performance one could expect. The other models should be compared based on their improvement from this baseline.

Next, we find that DistilBERT and RoBERTa models are 86% accurate, with similar numbers for recall, precision, and F1. AUC is 0.77 for RoBERTa and 0.76 for DistilBERT. The classical machine-learning baseline XgBoost is significantly worse in terms of AUC. Additional metrics and evaluation are reported in Appendix A1.

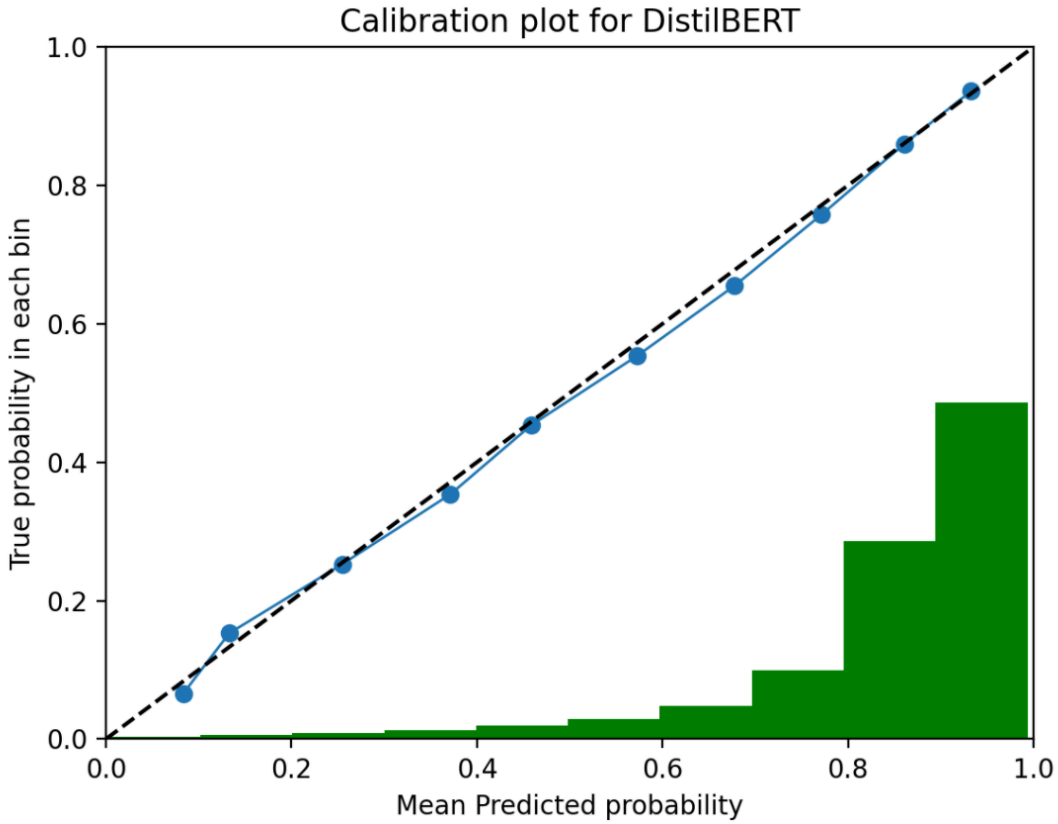
Table 3: Main Classification Metrics

Metrics	Guess Distinctive	XGBoost	DistilBERT	RoBERTa
Accuracy	0.84	0.84 (0.84, 0.84)	0.86 (0.85, 0.86)	0.86 (0.86, 0.86)
Balanced Accuracy	0.50	0.52 (0.52, 0.52)	0.60 (0.6, 0.6)	0.63 (0.63, 0.63)
Precision (weighted)	0.71	0.82 (0.82, 0.82)	0.83 (0.83, 0.83)	0.85 (0.85, 0.85)
F1 (weighted)	0.77	0.78 (0.78, 0.78)	0.82 (0.82, 0.82)	0.84 (0.84, 0.84)
AUC	0.50	0.67 (0.67, 0.67)	0.76 (0.75, 0.76)	0.77 (0.77, 0.77)

Note. Main classifier evaluation metrics for the distinctive outcome using a decision threshold of 0.5. Guess Distinctive has been calculated by guessing the modal class for every data point (every trademark is distinctive). We use weighted metrics to take into account the uneven distribution of mark distinctiveness in the dataset (84% of trademarks are distinctive, and 16% not distinctive). Balanced accuracy is calculated by taking the unweighted average recall across the two classes. The numbers in the bracket below represent a 95% confidence interval on the test set by bootstrapping 30 times with replacement.

Next, we want to check whether the models are well calibrated. In Figure 3, we divide the test dataset into ten bins by predicted outcome probability (probability that mark is distinctive) and compare them with the true probability in each bin. Ideally, the calibrated probability should be equal to the true probability for each bin. As can be seen in the figure, as the calibrated probabilities and true probabilities are very close for all the bins, we can say that the model is well calibrated, and can predict probabilities correctly. A calibrated probability of 0.9 implies that we can expect the mark to be distinctive 9 out of 10 times. While this is useful for decision-making, the graph also shows that the model is not that confident most of the time, with most of the probability mass being centered around the outcome base rate (84%). We provide the calibration plot for RoBERTa in the appendix (Figure A3).

Figure 3: Calibration Plot for DistilBERT



Note. The dashed black line shows the reference/ideal calibration. The x-axis gives bins of the predicted probabilities, and the y-axis (blue line) shows the true probability in each bin. The histogram depicts the distribution of predicted probabilities using 10 bins in percentage.

4.2 Varying Confidence in Predicting *Abercrombie*

As our model delivers 86% accuracy, with similar numbers for recall and precision, one may wonder how useful such a model may be for trademark law and practice. At first sight, 86% accuracy might seem nowhere close to approximating decisions by humans, and pretty much the same accuracy to guessing the modal category "distinctive" (84%). However, the performance metrics are comparable to other studies making predictions to support judicial decisions (such as Kleinberg et al. 2017, Ash et al. 2023, and Ash & Marangon 2023). Human decision-makers make errors, they render inconsistent decisions, and legal decisions get appealed and overturned on a frequent basis. Therefore, the accuracy of human decisions has important limitations as well.

More importantly, anecdotal evidence from trademark practice raises the question what the relevant benchmark of human decision-making is in such cases. As the Court of Appeals for the Second Circuit has acknowledged, placing a trademark on the *Abercrombie* spectrum "is far from an exact science, and [...] the differences between the classes, which is not always readily

apparent, makes placing a mark in its proper context and attaching to it one of the [*Abercrombie*] labels a tricky business at best" (Banff, Ltd. v. Federated Dep't Stores, Inc. 1988: 489). The Court of Appeals for the Seventh Circuit noted that distinguishing between suggestive and descriptive marks is difficult and is "often made on an intuitive basis rather than as the result of a logical analysis susceptible of articulation" (Union Carbide Corp. v. Every-Ready Inc. 1976: 379). And, in the words of the Court of Appeals for the Fifth Circuit: "The labels [of the *Abercrombie* spectrum] are more advisory than definitional, more like guidelines than pigeonholes. Not surprisingly, they are somewhat difficult to articulate and to apply" (Zatarain's, Inc. v. Oak Grove Smokehouse, Inc. 1983: 790). A leading treatise notes that delineating between descriptive and suggestive marks is "subjective" and "intuitive" (McCarthy 2023: §11.70). Perhaps as a result of these difficulties, Beebe (2006: 1635) finds that many district courts make little use of the *Abercrombie* spectrum in the context of likelihood of confusion inquiries (see also Fromer 2011: 1912-13).

As a result, it is not entirely clear how well humans will perform when deciding whether or not a trademark is inherently distinctive. It is unrealistic to expect 100% accuracy for humans, as they may make mistakes, may lack clear guidance from the vague *Abercrombie* spectrum test, or may just disagree with other humans in particular cases. An advantage of our approach is that our model provides predicted probabilities, which give a score of the difficulty of a case. This confidence score indicates how difficult it is for the machine to determine whether a trademark is distinctive or not. but might not reflect the difficulty for humans. Indeed, some cases that our model predicts easily might still be difficult for humans.

To explore the varying confidence in predicting a trademark's location on the *Abercrombie* spectrum, we deviate from a uniform decision threshold of 0.5 for all trademarks and analyze model precision of the calibrated probabilities at various decision thresholds. We simulate decisions based on nine decision thresholds, corresponding to predicted outcome probabilities. For a particular decision threshold (10%, 20%, ..., 90%), if the calibrated probability is above the decision threshold, we classify it as distinctive, otherwise as non-distinctive. Evaluating separate precision values for distinct and non-distinct marks at different decision thresholds provides us with a metric for how often the model is predicting correctly, conditional on the decision specified by that threshold.

Table 4 illustrates the precision for distinctive and non-distinctive marks at different decision thresholds for DistilBERT (see Appendix Figure A2 for a graphical representation). In this table, rows correspond to decision thresholds. That is, for a given threshold X (between 0.0 and 1.0), we assign trademarks as non-distinctive if the predicted probability Y-hat < X and distinctive if Y-hat > X. For example, for the standard decision threshold of 0.5, if the model predicted probability is between 0.0 and 0.5, we would assign non-distinctive for Y-hat < 0.5 and distinctive otherwise. The second and third columns indicate, for each threshold, the share of

observations that are below that threshold (second column) or above it (third column). The sum of the two columns adds to 100%. For example, only 0.31% of observations are assigned a Y-hat below 0.1 (first row, 2nd column), while 45.77% of observations are assigned a Y-hat above 0.9 (last row, 3rd column). The fourth and fifth columns indicate, for each decision threshold, the associated precision for each class (0 = non-distinctive, fourth column; 1 = distinctive, fifth column). For example, take the standard decision threshold of 0.5. In the set of predictions for class 0 (non-distinctive), the assignments are 66.3% accurate (4th column). In the set of predictions for class 1 (distinctive), the assignments are 86.72% accurate (5th column)

As Table 4 shows, the model becomes more precise for each class as we become more stringent in deciding for that class. For example, for a decision threshold of 0.8, a decision for the distinctive class (class 1) is accurate 90.39% of the time. Analogously, for a decision threshold of 0.2, the mark has an 88.29% chance of not being inherently distinctive. Overall, the model is particularly good at the tails of the distribution: for trademarks where the model predicts that the mark can be published with a very high (or very low) probability, the precision of this prediction is actually very high (>90%). In the appendix, we present a similar analysis for a fine-tuned RoBERTa model (Appendix Table A1, Figures A3 & A5).[8]

Table 4: Precision at Different Levels of Decision Threshold (DistilBERT)

Decision Threshold	% data < threshold	%data > threshold	Precision 0 < threshold	Precision 1 > threshold
0.1	0.31%	99.69%	91.66%	84.25%
0.2	0.96%	99.04%	88.29%	84.71%
0.3	1.87%	98.13%	81.59%	85.26%
0.4	3.14%	96.86%	73.99%	85.90%
0.5	5.09%	94.91%	66.39%	86.72%
0.6	8.04%	91.96%	57.72%	87.66%
0.7	13.01%	86.99%	47.98%	88.80%
0.8	23.45%	76.55%	36.80%	90.39%
0.9	54.23%	45.77%	23.82%	93.30%

[8] We run robustness checks in Appendix A.3 on whether the intuition holds that the more difficult it becomes to locate a trademark on the Abercrombie spectrum, the lower the model confidence is. As a rough proxy for how difficult it is to locate a trademark on the Abercrombie spectrum, we use the length of the trademark application procedure. Appendix A.3 shows that there is a negative relationship between length of procedure and model confidence.

Note. Column Precision 0 < threshold signifies the precision of mark not getting published if calibrated output probability is less than the corresponding decision threshold. Precision 1 > threshold signifies the precision of mark getting published if calibrated output probability is greater than the corresponding decision threshold.

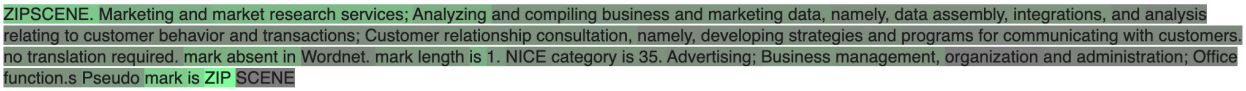
It is an interesting feature of our model that we can use these probabilities as assistance for deciding when to follow the classifier. If, for example, one feeds the model with a particular trademark and the model returns that the trademark has a 90% chance of being inherently distinctive, one also learns that the model's "confidence" in this prediction is very high (about 93%). If, however, the model returns for another trademark a 50% chance of being inherently distinctive, one also knows that the model is much less confident in this prediction (about 86%).

4.3 Explaining Model Predictions

We would like to understand which words in a trademark application are highly responsible for affecting our model's prediction towards distinctiveness on an instance as well as data level. To this extent, we explore separate instances and different Nice classes. To get intuition, we show example text documents with highlighting of the words according to how much they contribute to the model's prediction. We use a metric from the model explanation literature, called SHAP score (also known as Shapley value), which summarizes the contribution of an input feature to a model prediction (Lundberg & Lee, 2017; Molnar 2023).

For example, Figure 4 shows the textinput for the mark 'ZIPSCENE'.[9] Words that are highlighted in green have a higher SHAP score, indicating that they contribute to a finding of "is distinctive." The contribution of 'ZIP' outweighs the negative contribution of 'SCENE', thus resulting in the model's decision for distinctiveness (see also Figure A22). In Appendix 10.1, we expand this analysis beyond the single example of ZIPSCENE. We analyze SHAP scores for samples having different levels of predicted model probabilities of distinctiveness. We observe the influence of mark shifts from negative to positive as the probability of distinctiveness increases (Appendix 10.1).

Figure 4: Model Explanation results for ZIPSCENE



Note. Word Contribution Towards Distinctiveness for ZIPSCENE. P(distinctive)=0.91. Green implies a positive contribution toward distinctive; and dark gray implies a negative contribution toward not-distinctive. SHAP scores for BERT models are on a sub-word level. We aggregate the SHAP scores for a word to produce the figure above. "ZIP" and "SCENE" have high contributions, although in opposite directions.

[9] We use all the features mentioned in Table 2. Highlighting gives SHAP scores aggregated at word level. We remove the special tokens '[SEP]' and '[CLS]' so that they are not attributed to model probability through SHAP.

To further investigate the importance of mark names to effect model predictions, we investigate the influential words on a Nice class level using their SHAP scores. For this analysis, we focus on Nice classes 12 (vehicles), 16 (paper/cardboard), 32 (beverages), 43 (food/drinks), and 25 (clothing), as they provide well-defined categories of goods or services. Table 5 presents the top five words, their contributions, and associated marks for each of the Nice categories. From the table, it becomes clear that mark names are essential for determining the model's predictions, and are the most important feature. For the selected Nice classes, all but one of the top words contributing towards distinctiveness are directly associated with mark names.

Table 5: Top 5 words from NICE Classes 12, 16, 32, 43 and 25

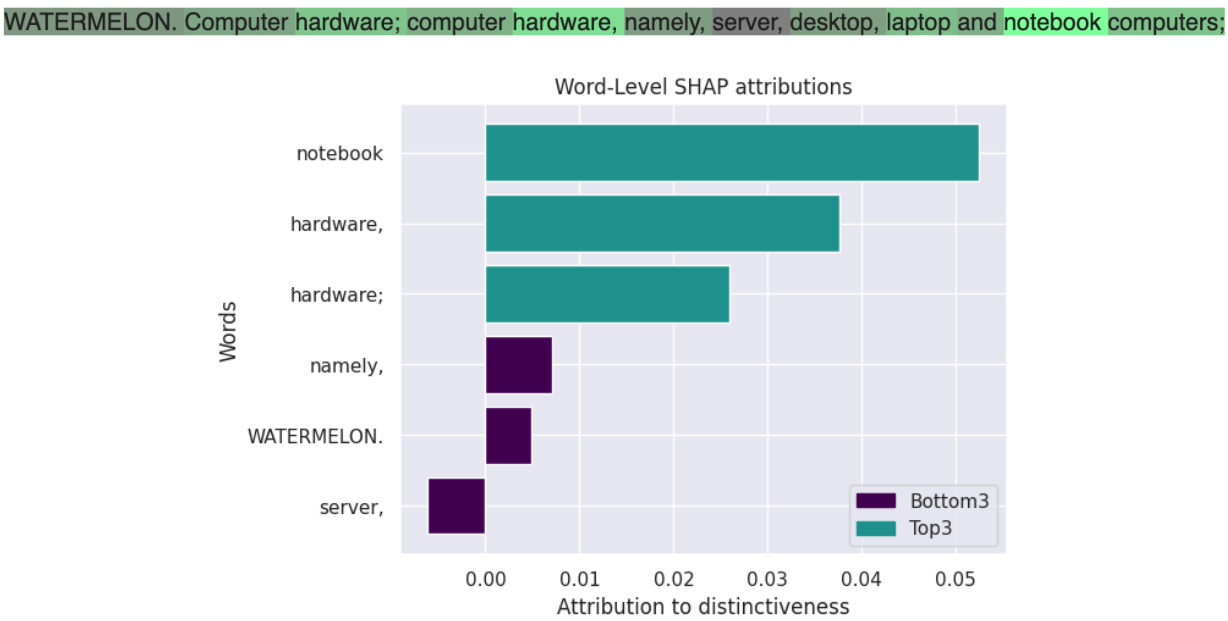
Nice Class	Words (Associated Marks)	SHAP Contribution (max value = 1)
12 (Vehicles)	VIVA (VIVA LA CROTCH!)	0.14
	III (BIN III)	0.14
	GET (GET UNSTUCK)	0.14
	A' (A' LA CART)	0.13
	CUSTOM (CUSTOM WHEEL HOUSE; CAYO CUSTOM BOATS)	0.12
16 (Paper / Cardboard)	360 (AWARENESS 360 SITUATION REPORT)	0.44
	= (REFLECTION + INSPECTION = NEW DIRECTION)	0.25
	MB(MB COUPON BOOK)	0.18
	JZ (JZ DESIGNS)	0.16
	WAD (BROCCOLI WAD)	0.16
32 (Beverages)	Z (HUPER Z)	0.15
	BSD (PLANT BSD)	0.14
	74 (BIOACTIVE WATER 74)	0.13
	TY (TY WINE ON)	0.13
	III (TIER III)	0.12
43 (Food / Drinks)	FK (FK YOUR DIET)	0.23
	ABC (ABC ARTISANAL BURGER COMPANY)	0.16
	MATE (YERBA MATE BAR)	0.15
	MIKE (SUSHI MIKE)	0.14
	NY (TONY'S NY PIZZA, NY NY PIZZA)	0.14
25 (Clothing)	connector (description of mark COUTURIETTE - a detachable connector for modifying a blouse or a shirt into a bodysuit)	0.12
	JSM (JSM Designs)	0.11
	SINGLE (SINGLE OVER SETTLING, SINGLE STITCH, SEXY & SINGLE, SINGLE TONIGHT, and 4 others)	0.11
	AMERICAN (GREY AMERICAN, AMERICAN DIESEL, THE AMERICAN TEE CO., AMERICAN HEART, and 58 others)	0.10
	1000 (WOLVERINE 1000 MILE, THE PROGRESSORS 1000)	0.10

Note. The table displays the top five words for the Nice classes 12, 16, 32, 43 and 25. The marks in the parentheses correspond to the mark in which these words were present. In some cases, the word might be present in more than one mark. Thus, for each Nice class in the table, we determine the average SHAP score of a word in the dataset and select the top five words. SHAP contribution displays the word's contribution towards distinctiveness. Class 12 represents Vehicles; apparatus for locomotion by land, air or water. Class 16 represents paper and cardboard; printed matter; bookbinding material; photographs; stationery and office requisites, except furniture; adhesives for stationery or household purposes; drawing materials and materials for artists; paintbrushes; instructional and teaching materials; plastic sheets, films and bags for wrapping and packaging; printers' type, printing blocks. Class 32 represents beers; non-alcoholic beverages; mineral and aerated waters; fruit beverages and fruit juices; syrups and other preparations for making non-alcoholic beverages. Class 43 represents services for providing food and drink; temporary accommodation. Class 25 represents clothing, footwear, headwear.

Although the contribution of mark names towards deciding distinctiveness is essential from the model's perspective, it is also important to note that the model can focus on other parts of the input text. To test the performance of our model in this regard, we explore whether the model probabilities change when the same mark has different descriptions (statement text). Ideally, if no relationship between mark name and description existed, an increase in model probabilities would be indicative of arbitrary, distinctive mark. If some relationship between mark name and description existed, model probabilities could indicate either a suggestive (distinctive) or a descriptive (non-distinctive) mark.

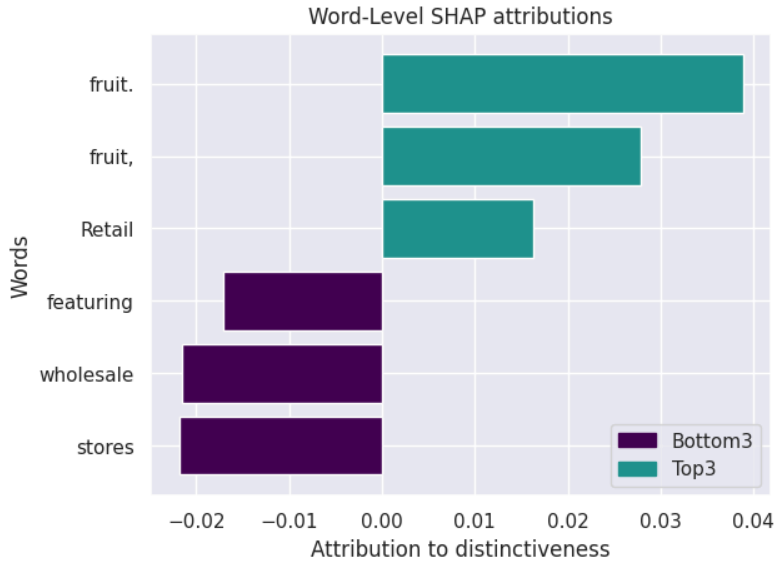
As an example, let us consider the mark "WATERMELON" and pair it with two different descriptions, one describing computer hardware and the other describing retail and wholesale fruit distributorship. Figure 5 presents our results. As expected, our model probabilities decrease from 0.67 in the context of computer services to 0.49 in the context of fruit distributorship. From the SHAP scores, we can see that words such as "stores" and "wholesale" have a high impact on the mark not being distinctive.

Figure 5: SHAP Scores for "Watermelon," Demonstrating Arbitrariness Dimension



(a) In the Context of Computer Hardware. Prob(distinctiveness) = 0.67

WATERMELON. Retail stores featuring fruit, wholesale distributorships featuring fruit, wholesale ordering services in the field of fruit.



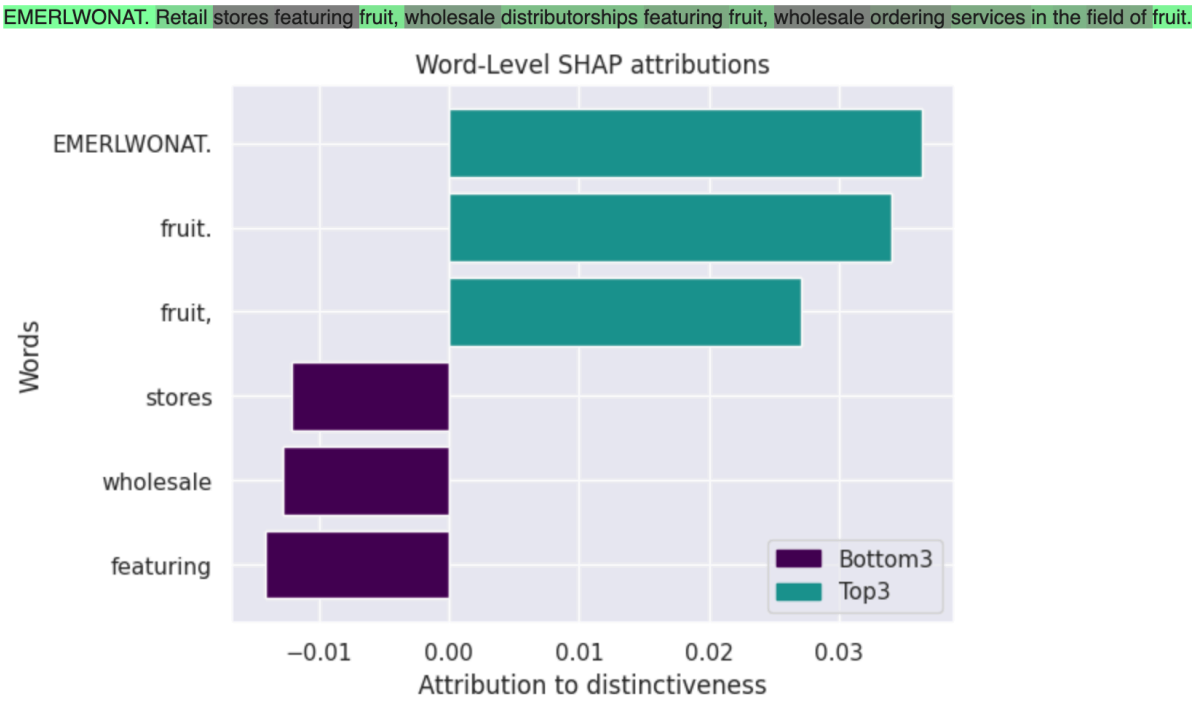
(b) In the Context of Fruit Distributorship. Prob(distinctiveness) = 0.49

Note. Green implies high contribution and gray implies low contribution towards distinctiveness. Changing the mark description changes the probability of distinctiveness. Note: Probability of distinctiveness changes with change in mark description. The barplot displays high negative attribution towards "featuring", "wholesale" and "stores"; resulting in lower probability of distinctiveness than "WATERMELON" in the context of computer hardware.

To extend our understanding of how SHAP can be used to understand our model's performance, we want to test how the model performs when giving made up words, as this can be an indicator of fanciful marks. We use the marks "EMERLWONAT", which is an anagram of "WATERMELON", and a made-up word "XADERMAC" paired with the fruit distributorship. In both cases, the model is able to predict the distinctiveness with high probabilities. Figure 6 and Figure A22 show the SHAP scores for these marks. In both cases, the mark names have a high impact on predicting distinctiveness.[10]

[10] Since "EMERLWONAT" is an anagram of "WATERMELON," it would be interesting to compare the attention distribution between "EMERLWONAT" and "WATERMELON." Figure A24 shows the attention score distribution for "EMERLWONAT." Compared to before, words attend to the mark name more strongly. Similarly, "fruit" attends to the mark name more strongly, but not as strongly as "computers." The average attention score of the mark name increased from 0.1082 ("WATERMELON") to 0.1207 ("EMERLWONAT").

Figure. 6: SHAP Scores for Invented Mark Term, Demonstrating Fancifulness Dimension



EMERLWONAT (Anagram of WATERMELON), in the Context of Fruit Distribution.
Prob(distinctive) = 0.66

Although the mark name is the most important feature, the model performance would be suboptimal if we just use the mark name as an input. Training the model only with marks produces an AUC of 0.69; 7 points below the model performance. Similarly, training the model without marks produces an AUC of 0.71, which is again lower than the model performance in Table 3. As a consequence, we think that both mark name and other features are important inputs when analyzing which features of a trademark application contribute to our model's prediction.

5. Implications

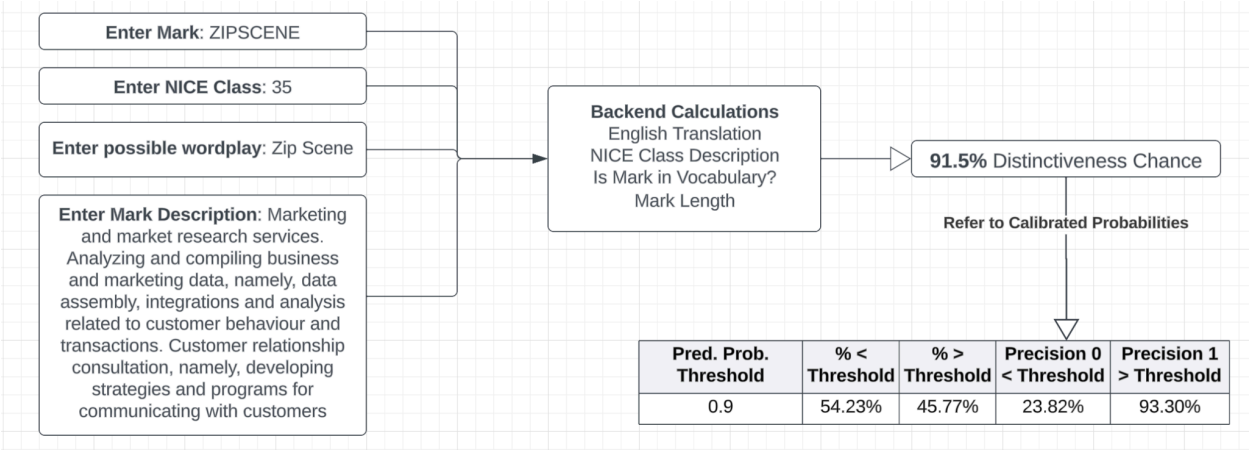
As Section 4 showed, our machine-learning model does not perform particularly well when applying a uniform decision threshold to all trademark applications. But in subsets of the data where the model is confident in its predictions, it performs astonishingly well – with over 90% accuracy. In this section, we explore the implications our model may have for trademark practice and theory. First, we outline the contours of a decision-support system based on our model that may pave a road towards a robot trademark clerk (not judge). Second, we discuss the implications the varying confidence in locating trademarks on the *Abercrombie* spectrum has on the *Abercrombie* spectrum itself.

5.1 Towards a Robot Trademark Clerk

The results from Table 4 show that our model performs reasonably well for trademarks with a high or low predicted probability. Hence, it could potentially serve as a backbone for a trademark decision-support system. This may not only be of interest to USPTO examining attorneys dealing with trademark applications and courts dealing with trademark infringement lawsuits. It could also be interesting for trademark attorneys and brand developers interested in assessing chances of registering a particular trademark.

Figure 7 shows a possible framework for such a decision-support system, illustrating the example of ZIPSCENE, a trademark that is part of our dataset (and which is registered as inherently distinctive). In this framework, users need to input the name of the trademark, its respective Nice classes, possible wordplays (for detecting pseudo marks), and a description of the trademark. The backend calculations involve text pre-processing, English translation of the trademark, and fetching the Nice class description. The model churns out the calibrated probability, which the user can refer to Table 4.

Figure 7: A Framework for a Decision-support System for the Trademark 'ZIPSCENE'



Pred. Prob. Threshold	% < Threshold	% > Threshold	Precision 0 < Threshold	Precision 1 > Threshold
0.9	54.23%	45.77%	23.82%	93.30%

Of course, we do not claim that our model is sufficient to replace a human decision-maker. As Section 6 demonstrates, we are aware of the many limitations a model such as ours would have to overcome before being able to replace a trademark specialist. But in our view, determining whether a trademark is distinctive or not is an excellent candidate for an automated decision-support system. Locating a trademark on the *Abercrombie* spectrum is a repetitive task that human experts have to perform nearly 800,000 times per year at the USPTO. Average trademark pendency has increased over the last years at the USPTO, and the agency has announced plans to hire an additional 140 trademark examining attorneys in 2023 and 2024 to speed up registration procedures (Fromer & McKenna 2024: 30). As trademark applications are a

high-volume business, trademark offices around the world have started to explore machine-learning techniques to speed up the trademark registration process (see Gangjee 2021: 184).

Once the decision-support system we envision has been set up, it would be easy to use and would give clear recommendations about which trademark application the human expert should focus his or her attention on. The system could also become an intuitive assistant to trademark examining attorneys, as the system would arguably develop similar background knowledge as trademark experts who have seen many trademark applications over a period of several years. We envision that the system would not only save trademark examining attorneys time; it could also increase the overall quality of their decisions. They would receive help from a machine to distinguish clear-cut cases, which are easy to decide and only require a quick look from the examining attorney, from cases that require more attention from the examining attorney. To assess the effectiveness of such a system, one could ideally test the system in a controlled policy experiment with trademark offices (for a policy experiment of the USPTO in patent law, see Pairolero et al. 2022).

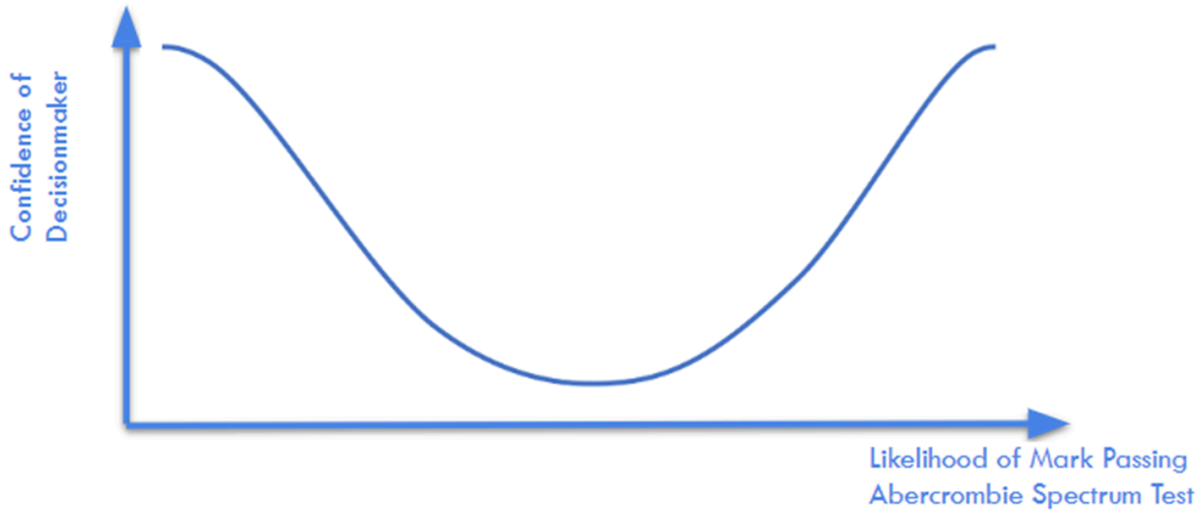
We want to add that such a trademark decision-support system would not only inform the trademark specialist about the likelihood that a trademark is inherently distinctive. As we have shown in Section 4.3, AI explanation tools such as SHAP can be used to analyze the features of a trademark application that drive our model's prediction. If a trademark decision-support system would become equipped with an explanation algorithm such as SHAP, this could point out to trademark examining attorneys and other decision-makers which factors lend the most weight to distinctiveness in the model, and would allow them to introspect on whether that coincides with their own experience and intuition. Furthermore, model explanations might help trademark scholars to develop a theory of the drivers of distinctiveness.

5.2 Overcoming *Abercrombie*

In our view, our study not only paves the road towards decision-support systems in trademark law. Perhaps more importantly, it also enables us to shed light on the limitations of the *Abercrombie* spectrum and the normative challenges resulting from these limitations.

A simplistic view on how the *Abercrombie* spectrum gets applied is that the human decision-maker learns about the definitions of the five tiers of the spectrum, gets experience by learning earlier decisions along the spectrum, and then applies the spectrum with the same confidence to all the different trademarks. However, such a view is naïve as the decision-maker will find it easier to determine the inherent distinctiveness of a trademark in clear-cut cases than in murky cases. In reality, the human decision-maker's confidence will be higher for both clearly distinctive and non-distinctive trademarks, whereas it will be lower for all the trademarks between these two extremes (see Figure 8).

Figure 8: Varying Confidence in Determining Trademark Distinctiveness



It is interesting that the performance of our machine-learning model roughly matches this conception of a human decision-maker's varying confidence in determining trademark distinctiveness. We argue that both human decision-makers and our model may perform reasonably well in clear-cut cases (where the probability of inherent trademark distinctiveness is either very high or very low), but that they perform much worse for middle-ground trademarks between these extremes.

While human decision-makers and our model may therefore share a similar varying confidence in determining trademark distinctiveness, they exhibit very different levels of transparency. As far as human decision-makers are concerned, their varying confidence in determining trademark distinctiveness is not easily observable. The USPTO examining attorney will not reveal how confident she was when deciding that a trademark is only descriptive rather than suggestive. And even though judges may sometimes indicate that they are less confident in their assessment of a trademark's distinctiveness, this information is neither provided in a systematic nor in a quantifiable or otherwise reliable manner.

By contrast, our model provides information about its own confidence jointly with its prediction. As described in Section 5.1, if one asks the model to predict whether a particular trademark is inherently distinctive, the model can assess its confidence in its own prediction. In our view, it is not only our model that is less confident in determining the distinctiveness of a murky, middle-ground trademark compared to clear-cut cases. Arguably, human decision-makers will have a similar variance in confidence. It seems intuitive that humans may perform well in cases where the model prediction is clear-cut (i.e., where the model has high confidence in whether a trademark is clearly distinctive or clearly not distinctive), while they may perform much worse in the murky areas between such clear cases. After all, our model was trained on actual decisions by

human decision-makers (USPTO examining attorneys). In this case, there may be an entire area of trademarks where both human decision-makers and machine-learning models perform poorly when deciding whether the trademark is inherently distinctive or not. These are the not-clear-cut cases, and the notion that determining trademark distinctiveness is a binary choice conceals the considerable variance decision-makers face when having to decide whether a middle-ground trademark is distinctive or not.

From our perspective, this raises the question whether the *Abercrombie* spectrum is the right framework for deciding whether murky, middle-ground signs should be protected as trademarks after all.[11] If decision-makers do not have high confidence in applying the *Abercrombie* spectrum test to such marks, perhaps one should actually refrain from the *Abercrombie* spectrum test. Rather, one could rely on other normative grounds when deciding whether such signs should be protected as trademarks or not. When it is not clear, for example, whether a sign is really inherently distinctive, perhaps trademark law should refrain from providing protection to such signs as overbroad protection may chill speech or unnecessarily clutter the linguistic space of potential trademarks (see Beebe & Fromer 2018; Ouellette 2014: 360). This could also push trademark owners towards more clearly arbitrary or fanciful marks (Tushnet 2017: 922). One could also rely on the conceptual distance between the primary meaning of a mark and the goods or services for which it is being used (Fromer 2022). Or one could activate a refined version of a secondary meaning test in such cases. Our paper is not the place to develop an alternative test of trademark distinctiveness. Yet, our empirical results cast some doubt on *Abercrombie*'s central role in trademark jurisprudence.

6. Discussion and Limitations

We have presented a machine-learning model that is trained on a semantic understanding of existing trademark applications in order to predict whether a trademark is inherently distinctive. Our machine-learning model looks at the trademark, its description and other features. It then tries to understand the meaning of the trademark. Arguably, our method mimics how an experienced USPTO examining attorney who is familiar with the USPTO decision practice would determine trademark distinctiveness.

In our view, using a natural-language model such as BERT that has learned the semantic relations among words and sentences is an important step towards determining trademark distinctiveness with quantitative approaches. Ten years ago, Ouellette (2014) proposed to determine trademark strength through conducting Google searches. She argued and demonstrated through case studies that the stronger a trademark is, the more top search results for the

[11] We are not the first to question whether the *Abercrombie* spectrum should be used to decide cases that are at the boundary between "descriptive or generic" and "arbitrary or fanciful" territory. For a discussion on whether suggestive marks should be treated as descriptive marks (i.e., requiring acquired secondary meaning) under the *Abercrombie* spectrum, see Linford (2015). Our focus in this paper is on the varying confidence of decision-makers in locating a trademark on the *Abercrombie* spectrum, depending on whether it is a clear-cut or a murky case. This problem may occur in all categories of the *Abercrombie* spectrum.

trademark will appear on Google. Courts have resisted such approaches on the ground that search results do not provide adequate information about the context in which a trademark is used (In Re Bayer Aktiengesellschaft (2007): 967; U.S. Patent and Trademark Office 2022b: §710.01(b); Gilson (2023): § 2.05[11]). Natural language processing tools take a step towards incorporating the semantic context of a trademark application.

Of course, our approach is not without limitations. First, as described in Section 3.1, while our approach allows us to predict whether or not a trademark is inherently distinctive, it does not allow us to locate the trademark on the five-tier *Abercrombie* spectrum in a fine-grained way. As trademark distinctiveness is a binary decision in trademark registration, our approach still delivers the relevant prediction in this administrative process. Second, our classifier is only trained with real-world trademark applications that have reached the USPTO in the past. We have not trained the classifier on trademark candidates that a company may have considered during its brand development process, but decided to never put forward to the USPTO. If such trademark candidates differed systematically from the observable trademark applications, we cannot exclude that our model would perform differently on such trademark candidates. However, we have tested our model with court decisions on trademark distinctiveness for out-of-sample predictions. Appendix A.2 reports results from this validation.

Third, while using BERT as a natural-language processing approach enables our study to understand the semantic features of trademark applications, choosing this machine-learning approach over others has implications for the interpretability of our results. With the emergence of large autoregressive language models such as GPT, one might want to use these models instead of BERT. The potential advantage of GPT over BERT is that it has access to an implicit general knowledge base and can reason over different factors. BERT, on the other hand, can be fine-tuned for our specific task and provides calibrated prediction probabilities. Hence, while GPT-type models open up exciting research avenues for trademark scholarship, the BERT approach is better-suited to our approach and aims. Our model is built specifically to follow USPTO decisions using information that is specified in the statute. GPT, meanwhile, would make a binary guess based on its general knowledge, rather than on what the USPTO has previously decided. In line with that, our model performs better at this task than recent baselines using GPT (e.g., Guha et al. 2022). The varying confidence in predicting *Abercrombie* (see Section 4.2) led us to challenge the notion that the *Abercrombie* spectrum is a good approach to decide about trademark registrability for murky, middle-ground trademarks (see Section 5.2). Using GPT for our task as in Goodhue & Wei (2023) or Guha et al. (2022) would not produce the relevant model precision metrics that are necessary to engage in such normative debate. In addition, using a model such as BERT enables us to use AI explanation tools to explore why our model makes particular predictions.

Fourth, it should be noted that the model's confidence scores are concentrated in the middle of the distribution. This means that the model is not confident for most of the trademarks, in which case the system would not be very useful for the decision-maker. Given that, for example, only 2.5% of the data points get a confidence score over 90%, that puts limits on the system's performance. Note that the low confidence could be due to subjectivity, noise, and other irreducible errors in the USPTO process. In that case, no model would be able to confidently predict large swaths of the data.

Fifth, fine-tuning models and deriving explanations on the entire dataset can be computationally expensive. In our study, we used Nvidia V-100 GPUs to fine-tune a comparatively small language model; each epoch taking approximately 2 hours. For larger models such as RoBERTa, this time increased to approximately 3.5 hours. Further, calculating SHAP values on an instance level is quick, but to have a global view of the features in the dataset is computationally challenging. As an example, for Nice class 12, which had only 3890 examples in our test dataset, it took us approximately 2 hours. For Nice classes having multiple thousands of examples across the years, this approach can be computationally infeasible.

Finally, our framework does not capture the trademark registration process in its entire complexity. While our system arguably mimics how an experienced USPTO examining attorney would determine trademark distinctiveness, it takes distinctiveness of trademarks as determined by the USPTO in the training set as a given (see Lim 2022: 1358-1359). Determining ground truth is a recurring problem for trademark distinctiveness, and our paper presents two approaches (USPTO decisions in the main analysis and court decisions in Appendix A2) to tackle this problem. For a project that aims at presenting a decision-support system and exploring the foundations of the *Abercrombie* test, these approaches seem the right starting points. Also, our model only covers word marks. An interesting extension of this work would be to expand the analysis to special form trademarks (including design elements or colors, as in the case of logos) and image trademarks by either translating images into text descriptors or by applying multimodal machine-learning techniques that work directly on both text and images. Furthermore, current trademark doctrine has many subtle rules for special cases, such as double entendre, composites, misspelling, foreign equivalents, or trademarks with similar pronunciation, which our current framework does not capture (see Goodhue & Xing 2023). Relatedly, if trademark examination was increasingly automated, and if brand development and trademark applications became increasingly automated as well, one could envision a future in which machines apply for trademarks which are then checked by machines. The recent trend to apply to "nonsense marks" (see Fromer & McKenna 2024) due to the particular design of the Amazon Brand Registry may point towards such a future. We think that it is for trademark law and procedure to provide a framework that minimizes negative effects of dubious trademark registration strategies (see Fromer & McKenna 2024: 67-68). A machine-learning model that makes predictions based on earlier decisions by trademark examining attorneys cannot solve

larger policy problems that the current trademark system faces. Yet, such a model could be used to identify potential candidates for nonsense trademark applications in a more refined way than by counting the number of consonants or vowels that appear in a word in a row (see Fromer & McKenna 2024: 41). Thereby, machine-learning models could become part of the solution to nonsense marks. Overall, we view our framework not as a system that is ready to be used in the real world, but as a prototype that enables us to explore the extent to which the *Abercrombie* spectrum can be automated and what kind of normative tradeoffs such automatization would entail.

7. Conclusion

We started our journey with a vision to automate the *Abercrombie* spectrum in trademark law. Trademark scholars are often skeptical of such visions, as it is unclear whether artificial intelligence technologies will ever be able to reflect the holistic and human-centric approaches on which many complex trademark doctrines are built (Moerland & Freitas 2021: 284). If trademark doctrines such as inherent distinctiveness require a subjective and nuanced evaluation (see Moerland & Freitas 2021: 288; Katyal & Kesari 2020: 526; Gangjee 2021: 190; Lim 2022: 1362-1364), it is not clear whether we will ever be able to turn trademark procedure over to the machines.

We view our study as an endeavor to discuss these challenges not in the abstract, but through an engineering approach of trial and error. We have presented a prototype for a robot trademark clerk and suggested trying out such a clerk in the field under controlled conditions. In this process, we have observed that our framework may provide helpful empirical input to discussions trademark scholars have only conducted on a theoretical level so far. In their discussion of proximity and distance costs, for example, Hemel & Ouellette (2021) provide a conceptual framework to think about the implications of trademark distinctiveness. But they lack an empirical exploration validating their framework. Our study provides a roadmap on how to quantify distance costs. In our baseline Xgboost model (see Appendix A.7), for example, we use the distance between text embeddings of the trademark and its description as a predictive feature. Furthermore, our efforts on how to infuse the *Abercrombie* spectrum – which is a highly nuanced, complicated trademark doctrine that may be applied in an inconsistent and subjective manner – with current machine-learning methods may indicate how challenging it is to define proper benchmarks for large language models (see Guha et al. 2023).

More importantly, our framework has raised doubts whether the *Abercrombie* spectrum should be the guiding principle for trademark registrability when trademark distinctiveness is uncertain, where the model-predicted probabilities are close to 50%. Our model performs poorly on these middle-ground trademarks, and we expect that this mimics the low confidence that human trademark experts have when deciding such cases. The *Abercrombie* spectrum has always focused on how consumers perceive a particular sign and then provided normative prescriptions

on how such signs should be treated by trademark decision-makers. But when trademark decision-makers have low confidence on how consumers perceive a range of signs, it is perhaps a mistake to let the *Abercrombie* spectrum derive normative prescriptions from a reality of sign usage that is murky and blurred.

The *Abercrombie* spectrum serves ultimately only as a heuristic to aid trademark examiners, lawyers, and judges in quickly determining whether, as a matter of trademark policy, a mark merits protection without any showing of acquired distinctiveness. It substitutes for an involved cost-benefit analysis of whether granting protection would benefit competition and consumers. When it was first formulated, the *Abercrombie* test was recognized as a means of reducing this analysis to what were often more tractable and easily-answered questions: Is the mark a new, coined term? Does the mark have any semantic relation to its products or services? But when those questions are not easily answered, the *Abercrombie* test may do more harm than good. By focusing on the formal, semantic classification of a mark, it may distract from the core policy question of whether it is optimal to grant exclusive rights in a mark without any additional showing of acquired distinctiveness. In such cases, it may make sense to set the heuristic aside and return to and confront directly the underlying policy question otherwise latent in the *Abercrombie* test. By demonstrating the substantial limitations of the *Abercrombie* heuristic, machine-learning methods may teach us to separate those situations in which machine-learning models can effectively do the work of the *Abercrombie* test and of humans applying it from those situations in which the test should simply not be applied by humans or machines. Automating *Abercrombie* would therefore not turn trademark procedure over to the machines, but partially free trademark procedure from a test that partially produces inconclusive results.

References

Abercrombie & Fitch. Abercrombie & Fitch Co. v. Hunting World, Inc., 537 F.2d 4 (2nd Cir. 1976).

Ashtor, J. H. (2022). Modeling Patent Clarity. Research Policy 51, 1-15.

Banff, Ltd. v. Federated Dep't Stores, Inc. (1988). Banff, Ltd. v. Federated Dep't Stores, Inc., 841 F.2d 486 (2d. Cir. 1988).

Beebe, B. (2004). The Semiotic Analysis of Trademark Law. UCLA Law Review 51, 621-704.

Beebe, B. (2006). An Empirical Study of the Multifactor Tests for Trademark Infringement. In California Law Review 94, 1581-1654.

Beebe, B., & Fromer, J. (2018). Are We Running Out of Trademarks? An Empirical Study of Trademark Depletion and Congestion. Harvard Law Review 131, 945-1045.

Beebe, B. & Fromer, J. (2020). Patent and Trademark Office Trademark Office Actions Dataset, 2003–2019 (2020).

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System Proceedings of the 2nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785-794. https://doi.org/10.1145/2939672.2939785.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 1, 4171-4186. https://arxiv.org/abs/1810.04805.

European Union Intellectual Property Office (2023). Nice Class Headings. http://euipo.europa.eu/ec2/classheadings.

Fromer, J. (2011). The Role of Creativity in Trademark Law. Notre Dame Law Review 86, 1885-1920.

Fromer, J. (2022). Against Secondary Meaning. Notre Dame Law Review 98, 211-265.

Fromer, J. & McKenna, M. (2024). Amazon's Quiet Overhaul of the Trademark System. Forthcoming in the California Law Review. https://ssrn.com/abstract=4870984.

Gangjee, D. (2021). Eye, Robot: Artificial Intelligence and Trade Mark Registers. In N. Bruun et al. (eds.), Transition and Coherence in Intellectual Property Law: Essays in Honour of Annette. Cambridge University Press, 174-190.

Gilson, A., & Gilson, J. (2023). Gilson on Trademarks. LexisNexis.

Goodhue, J., & Wei, Y. (2023). Classification of Trademark Distinctiveness using OpenAI GPT 3.5 Model. https://ssrn.com/abstract=4351998.

Goodhue, J. & Xing, L. (2023). Addressing "Special Issues" in Classifying Trademark Distinctiveness Using GPT-3. https://ssrn.com/abstract=4582171.

Guha, N., Ho, D. E., Nyarko, J., & Ré, C. (2022). LegalBench: Prototyping a Collaborative Benchmark for Legal Reasoning. https://arxiv.org/abs/2209.06120v1.

Guha, N. et al. (2023). LegalBench: A Collaboratively Built Benchmark for Measuring Legal Reasoning in Large Language Models. https://arxiv.org/abs/2308.11462.

Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On Calibration of Modern Neural Networks 34th International Conference on Machine Learning, 1321-1330. https://dl.acm.org/doi/10.5555/3305381.3305518.

Graham, S., Hancock, G., Marco, A., & Myers, A. (2013). The USPTO Trademark Case Files Dataset: Descriptions, Lessons, and Insights. https://ssrn.com/abstract=2188621.

Hain, D. S., Jurowetzki, R., Buchmann, T., & Wolf, P. (2022). A Text-embedding-based Approach to Measuring Patent-to-patent Technological Similarity. Technological Forecasting and Social Change, 177, 121559.

Hemel, D., & Ouellette, L. (2021). Trademark Law Pluralism. University of Chicago Law Review 88, 1025-1080.

In Re Bayer Aktiengesellschaft (2007). In Re Bayer Aktiengesellschaft, 488 F.3d 960 (Fed. Cir. 2007).

Katyal, S., & Kesari, A. (2020). Trademark Search, Artificial Intelligence, and the Role of the Private Sector. Berkeley Technology Law Journal 35, 501-588.

Clark, K., Khandelwal, U., Levy, O. & Manning, C. (2019). What Does BERT Look At? An Analysis of BERT's Attention. In Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, pages 276–286, Florence, Italy.

LegalBench (2023). LegalBench: Task Overview. https://github.com/HazyResearch/legalbench/tree/main/tasks.

Lim, D. (2022). Trademark Confusion Revealed: An Empirical Analysis. 71 American University Law Review 1285-1365.

Lim, D., & Abbott, R. (2022). Computational Trademark Infringement and Adjudication. In R. Abbott (ed.), Research Handbook on Intellectual Property and Artificial Intelligence. Edward Elgar Publishing, 259-289.

Linford, J. (2015). The False Dichotomy Between Suggestive and Descriptive Trademarks. Ohio State Law Journal 76, 1367-1421.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L. & Stoyanov, V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. https://arxiv.org/abs/1907.11692.

Lundberg, S., & Lee, S. (2017). A Unified Approach to Interpreting Model Predictions. Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17), 4768-4777.

McCarthy, J. (2023). McCarthy on Trademarks and Unfair Competition, 5th edition. Thomson West.

Moerland, A., & Freitas, C. (2021). Artificial Intelligence and Trademark Assessment. In J.-A. Lee, R. Hilty, & K.-C. Liu (eds.), Artificial Intelligence and Intellectual Property. Oxford University Press, 266-291.

Molnar, C. (2023). Interpretable Machine Learning: A Guide for Making Black Box Models Explainable. https://christophm.github.io/interpretable-ml-book/.

Ouellette, L. (2014). The Google Shortcut to Trademark Law. California Law Review 102, 351-407.

Pairolero, N., Toole, A., Pappas, P., de Grazia, C. & Teodorescu, M. (2022). Closing the Gender Gap in Patenting: Evidence from a Randomized Control Trial at the USPTO. https://ssrn.com/abstract=4265093.

Princeton University (2023). WordNet: A Lexical Database for English. https://wordnet.princeton.edu.

Sanh, V., Debut, L., Chaumon, J., & Wolf, T. (2020). DistilBERT, a Distilled Version of BERT: Smaller, Faster, Cheaper and Lighter. https://arxiv.org/abs/1910.01108.

Shackell, C., & De Vine, L. (2022). Quantifying the Genericness of Trademarks Using Natural Language Processing: An Introduction with Suggested Metrics. Artificial Intelligence and Law 30, 199-220.

Showkatramani, G., Khatri, N., Landicho, A., & Layog, D. (2019). Deep Learning Approach to Trademark International Class Identification. Proceedings of the 18th IEEE International Conference on Machine Learning and Applications (ICMLA 2019), 608-612.

Trappey, C., Trappey, A., & Liu, B.-H. (2020). Identify Trademark Legal Case Precedents: Using Machine Learning to Enable Semantic Analysis of Judgments. World Patent Information 62, 101980.

Tushnet, R. (2017). Registering Disagreement: Registration in Modern American Trademark Law. Harvard Law Review 130, 867-941.

Union Carbide Corp. v. Every-Ready Inc. (1976). Union Carbide Corp. v. Every-Ready Inc., 531 F.2d 366 (7th Cir. 1976).

U.S. Patent and Trademark Office (2022a). FY 2022 Workload Table 16. https://www.uspto.gov/sites/default/files/documents/USPTOFY22WorkloadTables.xlsx.

U.S. Patent and Trademark Office (2022b). Trademark Manual of Examining Procedure (TMEP). https://tmep.uspto.gov/RDMS/TMEP/current.

U.S. Patent and Trademark Office (2023). Current Trademark Processing Wait Times. https://www.uspto.gov/dashboard/trademarks/application-timeline.html.

Vig, J. (2019). Visualizing Attention in Transformer-Based Language Representation Models. https://arxiv.org/abs/1904.02679.

Zatarain's, Inc. v. Oak Grove Smokehouse, Inc. (1983). Zatarain's, Inc. v. Oak Grove Smokehouse, Inc., 698 F.2d 786 (5th Cir. 1983).

Appendix A: Additional Analyses

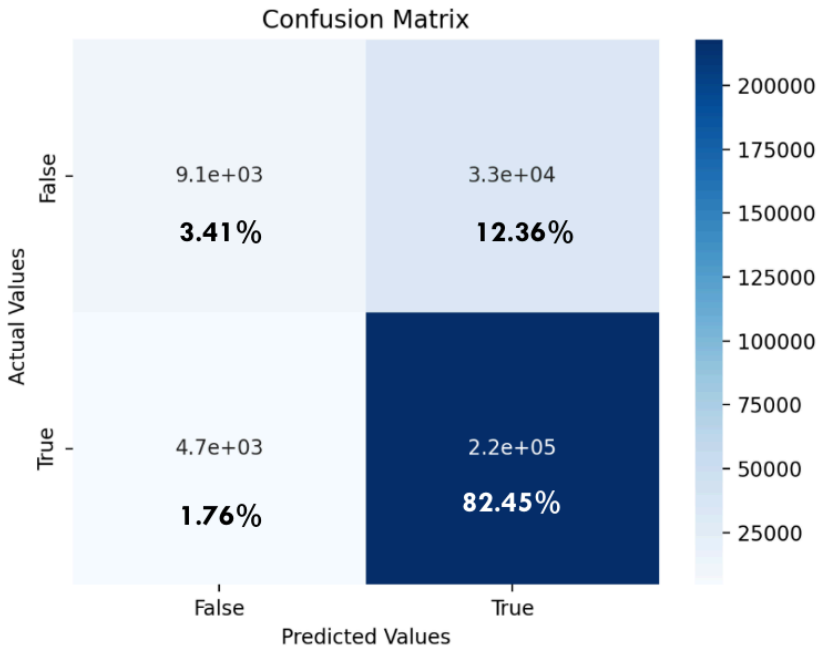
A.1 Additional Analyses for Distinctiveness Indicator

The following section presents additional results for the distinctiveness indicator. Figure A1(a) shows the confidence matrix of our DistilBERT classifier with a decision threshold of 0.5. Figure A1(b) shows the ROC curve and model AUC. Figure A2 represents a graphical interpretation of Table 4 (precision of our model for predicting distinctiveness indicator under various thresholds).

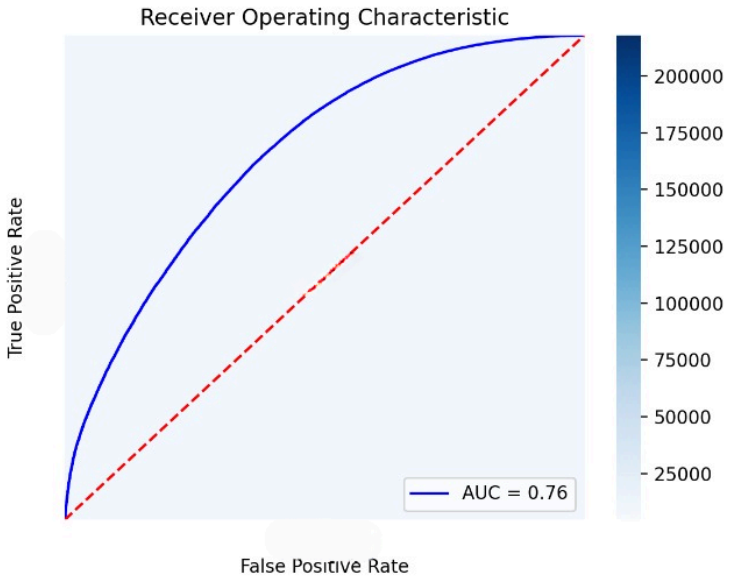
Further, we look at the results from the RoBERTa model. Figure A3 shows the calibration plot of our fine-tuned RoBERTa model, and Table A1 shows the precision of the RoBERTa model on varying decision thresholds. Similar to DistilBERT, we report the confusion matrix, ROC-AUC and a graphical representation of precision on varying thresholds in Figures A4 and A5. Lastly, we perform ablation studies on the inputs and report the classification metrics on varying inputs. We use only DistilBERT to perform the ablation study, as this process is computationally extensive.

Figure A1: **(a)** shows the confusion matrix with a decision threshold of 0.5, and **(b)** shows the AUC of 0.76 achieved by the fine-tuned DistilBERT model.

(a)

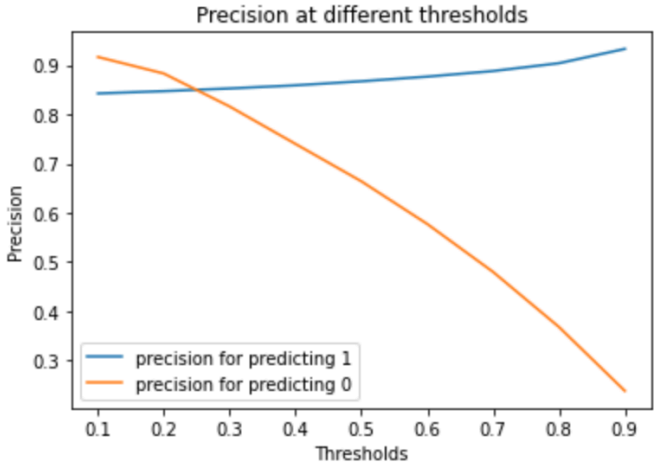


(b)



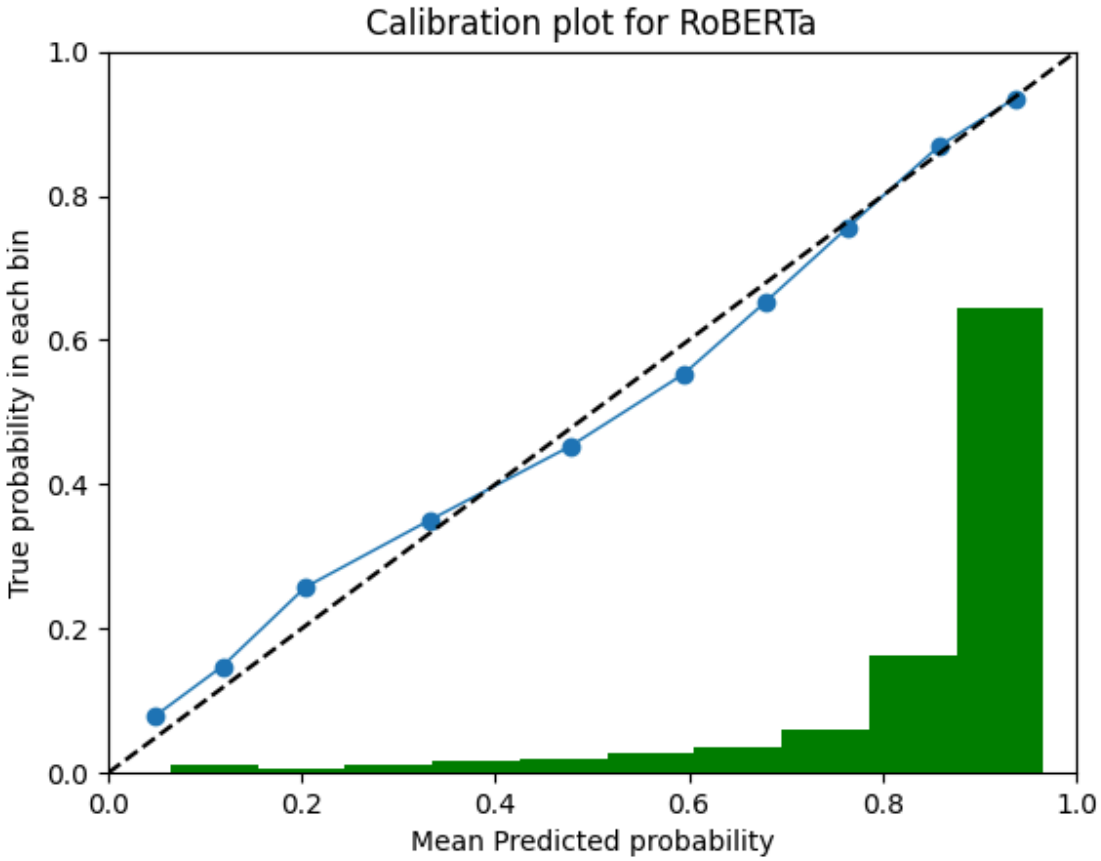
Note. Figure A1 is a graphical representation for Table 3.

Figure A2: Graphical Representation of Precision for Distinctiveness (Predicting 1) and Non-distinctiveness (predicting 0) of Marks for Varying Decision Thresholds (Fine-tuned DistilBERT)



Note. This is a graphical representation for Table 4.

Figure A3: Calibration Plot for RoBERTa



Note. The dashed black line shows the reference/ideal calibration of a model. X-axis shows the predicted probabilities, and Y-axis (blue line) shows the true probability in each bin. The histogram depicts the distribution of predicted probabilities using 10 bins in percentage.

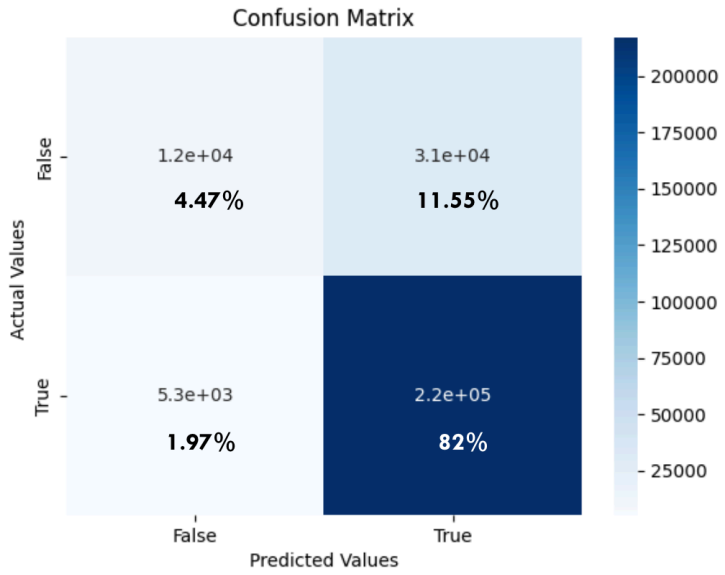
Table A1: Precision at Different Levels of Decision Thresholds (RoBERTa)

Decision Threshold	% data < threshold	% data > threshold	Precision 0 < threshold	Precision 1 > threshold
0.1	0.70%	99.30%	95.14%	84.57%
0.2	1.37%	98.63%	91.73%	85.07%
0.3	2.38%	97.62%	86.58%	85.74%
0.4	4.09%	95.91%	78.32%	86.67%
0.5	6.14%	93.86%	69.63%	87.52%
0.6	9.10%	90.90%	60.16%	88.44%
0.7	13.17%	86.83%	51.49%	89.40%
0.8	20.46%	79.54%	41.54%	90.59%
0.9	51.83%	48.17%	24.97%	93.68%

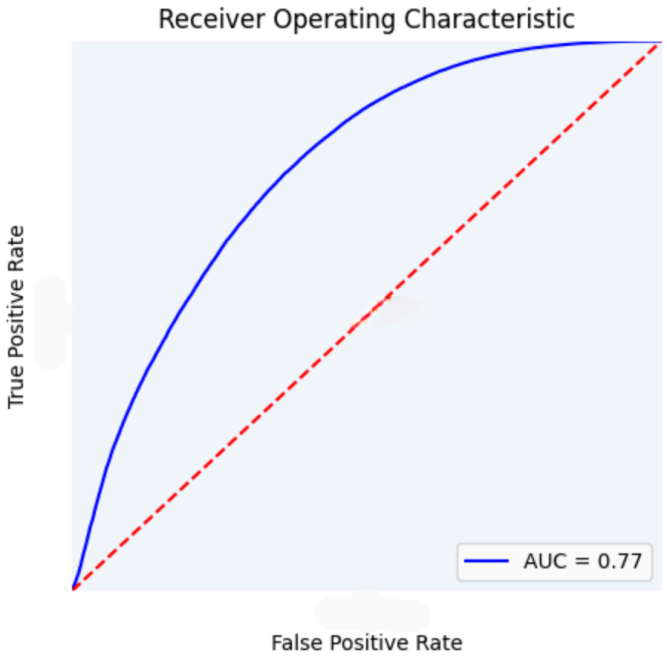
Note. Column Precision 0 < threshold signifies the precision of mark not getting published if calibrated output probability is less than the corresponding decision threshold. Precision 1 > threshold signifies the precision of mark being distinctive if calibrated output probability is greater than the corresponding decision threshold.

Figure A4: **Confusion Matrix and ROC Curve**

(a)

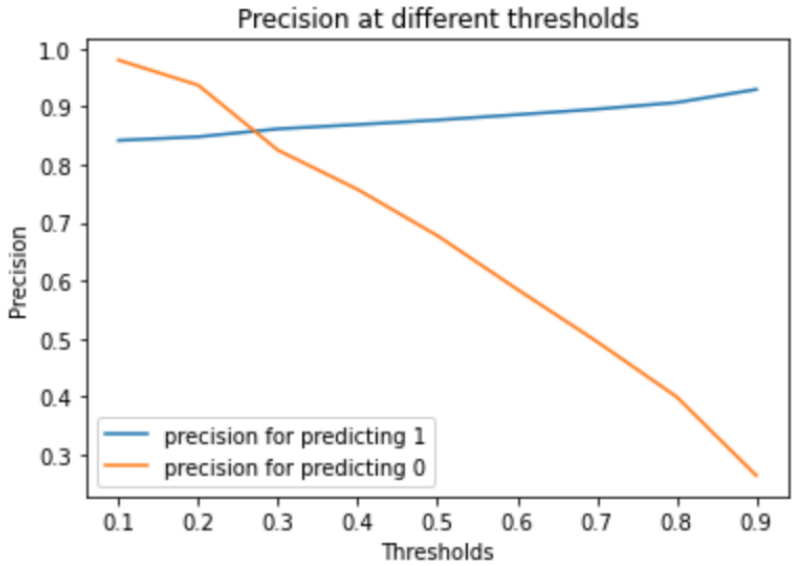


(b)



Note. Panel (a) shows the confusion matrix with a decision threshold of 0.5, and (b) shows the AUC of 0.77 achieved by the fine-tuned RoBERTa model.

Figure A5: Model Precision at Different Decision Thresholds



Note. A graphical representation of precision for distinctiveness (predicting 1) and non-distinctiveness (predicting 0) of marks for varying decision thresholds (fine-tuned RoBERTa).

Table A2: Performance Metrics for Minority Class

Metrics	Guess Distinctive	XGBoost	DistilBERT	RoBERTa
Recall (minority)	0	0.04	0.21	0.27
Precision (minority)	-	0.69	0.66	0.69
F1 (minority)	0	0.08	0.32	0.39

Note. As the dataset is imbalanced, we also provide metrics for the minority class (not distinctive).

Table A3: Classification Metrics for Model Variants

Metrics	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Accuracy	0.85	0.86	0.86	0.86	0.86	0.86	0.86	0.86
Recall (weighted)	0.85	0.86	0.86	0.86	0.86	0.86	0.86	0.86
Precision (weighted)	0.82	0.84	0.84	0.84	0.84	0.84	0.84	0.83
F1 (weighted)	0.80	.82	.82	.82	.83	.82	.82	0.82
AUC	.69	.76	.76	.76	.76	.76	.76	.76
Mark	X	X	X	X	X	X	X	X
Statement		X	X	X	X	X	X	X
Translated Mark			X	X	X	X	X	X
Word in Dictionary				X	X	X	X	X
Mark Length					X	X	X	X
NICE Class						X	X	X
Pseudo Mark							X	X
NICE Description								X

Note. DistilBERT classifier evaluation metrics for the distinctiveness outcome using a decision threshold of 0.5. We use weighted metrics to take into account the uneven distribution of inherently distinctive marks in the dataset (84% of marks are inherently distinct, and 16% are not).

A.2 Model Application to Court Decisions

In order to further validate the performance of our model, we have provided our model with out-of-sample trademarks where courts have ruled on their location on the *Abercrombie* spectrum, thereby serving as an alternative test set. We have identified 965 federal court decisions in which courts ruled on whether a particular trademark is generic, descriptive, suggestive, arbitrary, or fanciful. In many cases, the trademark registration number, Nice class, and statement text are not available from the court decision. We used research assistants to map these court decisions to trademark registrations, thereby retrieving the necessary supplementary information. We then coded the outcome of the court decision with regard to where a trademark is located on the *Abercrombie* spectrum. To ensure that our model has not seen the respective trademark in the training, validation, or test stage, we removed court decisions dealing with trademarks from our training, validation, and test datasets. This left us with 516 federal court decisions.

This section reports results on the model capacity by applying it to these court decisions. Table A4 shows the data distribution. The data distribution presented in this dataset is significantly different from our USPTO training dataset: while the USPTO training dataset had about 84% distinctive marks, our current court dataset has about 50% distinctive marks. Since the data distribution is skewed, we test our model by balancing the data by mark types, where each mark is allocated an equal number of samples. As generic marks have the lowest count, we get 42 samples for all five types of marks.

Table A5 shows the model's performance on this dataset across the different balancing strategies. Despite the significant shift (approximately 34%) in data distribution, our model loses only 16 points in AUC compared to Table 3 when utilizing the complete dataset. Our AUC loses only 12 points when we balance the dataset based on mark types. In general, the robustness of AUC values show that the model is able to generalize well. However, since the dataset is skewed towards non-distinctiveness, we adjust our threshold from 0.5 to 0.69 by optimizing the difference between True Positive Rate and False Positive Rate from the AUC Curve. We follow the same strategy for balanced datasets. The reason for choosing different thresholds, even for balanced dataset, is that our original dataset was skewed towards the distinctiveness indicator.

We then assess the model's performance in each of the five categories on the *Abercrombie* spectrum. Figure A6 shows the 95% confidence interval of probabilities across different mark types. The figure suggests that the model is able to capture the rising level of distinctiveness through the predicted probabilities, despite a significant change in the distribution.

Table A4: (a) Spectrum Distribution, and (b) Distinctiveness Distribution

(a)

Mark Type	Count
Generic	42
Descriptive	225
Suggestive	130
Arbitrary	73
Fanciful	46
Total	516

(b)

Mark Type	Percentage
Distinctive	50.39%
Non-distinctive	49.61%

Note. Data distribution for court cases (a) according to the *Abercrombie* spectrum, and (b) according to distinctiveness.

Table A5: Model Performance on Court Decisions

(a)

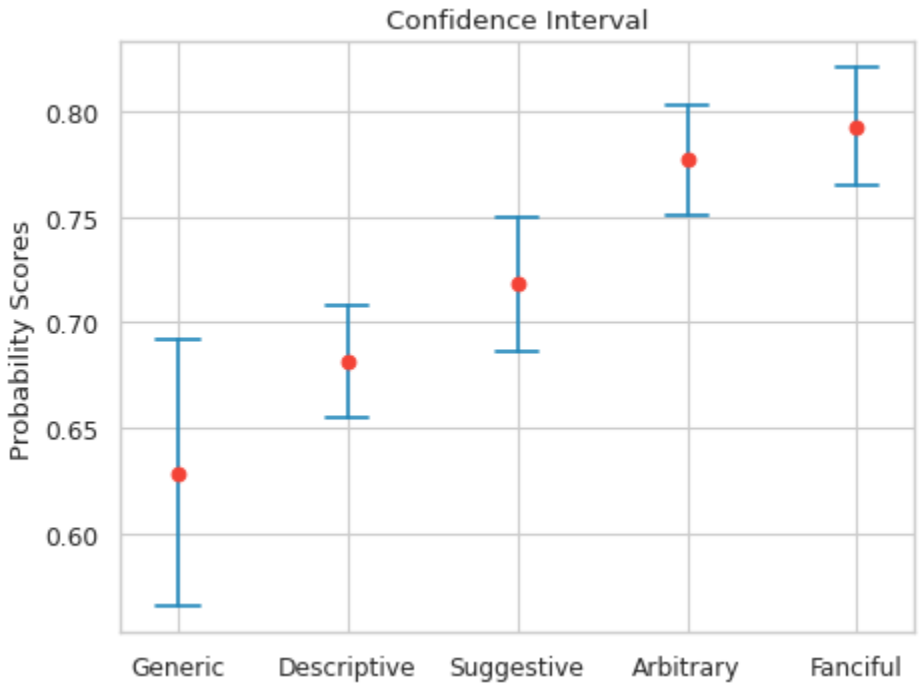
Metric	Original Dataset	Balanced by Mark Type
Accuracy	0.58	0.63
Precision (weighted)	0.60	0.64
Recall (weighted)	0.58	0.63
F1 Score (weighted)	0.57	0.63
AUC	0.60	0.64

(b)

Metric	Original Dataset	Balanced by Mark Type
Accuracy	0.54	0.63
Precision (weighted)	0.63	0.65
Recall (weighted)	0.54	0.64
F1 Score (weighted)	0.47	0.58
AUC	0.60	0.64

Note. (a) shows the model performance across different strategies using an optimal threshold. We use a threshold of 0.69 for the original dataset and 0.73 when balancing by mark type. All the thresholds were found by optimizing the difference between True Positive Rate and False Positive Rate from the AUC curve. (b) shows the model performance across different strategies when using 0.5 as the decision threshold.

Figure A6: Predicted Probabilities By True Mark Type in Court Data



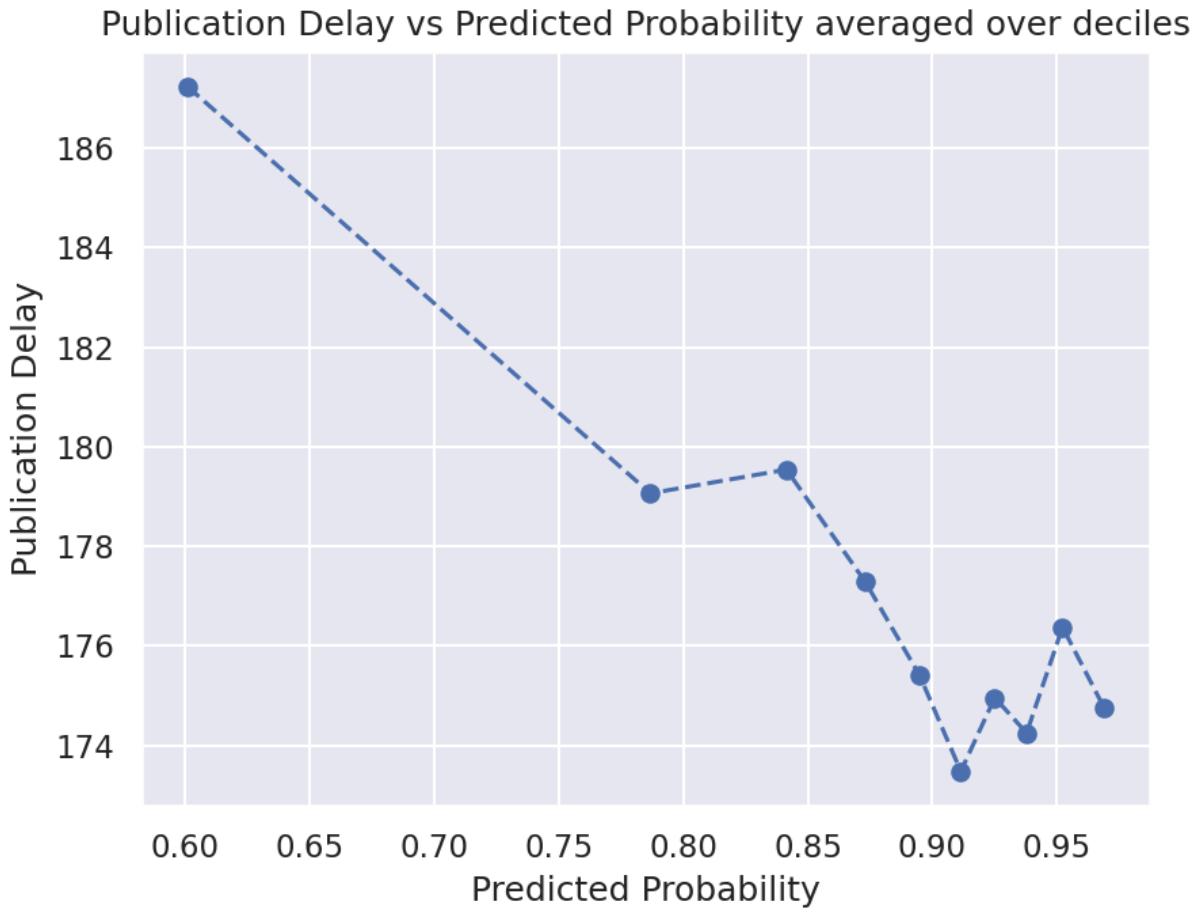
Note. The red dot represents the average predicted probability for each mark type for the entire dataset. As the distinctiveness increases across the spectra, the average model probability increases as well.

A.3 Lower Model Confidence is Associated with More Delayed Publication

The model's confidence in its predictions, proxied by the outputted predicted probability given an inputted data point, can be interpreted as measuring the "difficulty" of determining the distinctiveness of the associated mark. That is, if the model is 99.9% confident, that is likely to be an "easy" case and the trademark examining attorney should be able to decide it quickly. On the other hand, if the model is less confident, say 55% predicted probability, that would tend to be a "difficult" case, and the trademark examining attorney would likely have to do more investigation or otherwise undertake a longer decision process.

If this reasoning holds, we would observe in the data that the model should be less confident – that is, the predicted probability should be closer to 50% – when the trademark office takes longer to publish a mark. Limiting the sample to published marks, we construct the publication delay as the difference in days between publication date and filing date. In Figure A7, we plot the predicted probability against the publication delay. To construct this plot, we divide the dataset into 10 deciles of predicted probabilities and then calculate the average predicted publication delay in each decile. Consistent with the above reasoning, we see a negative relationship of predicted probabilities with publication delay. One should be careful, however, in interpreting this finding. Delays in trademark procedures can have multiple reasons, not all of which have to do with the perceived difficulty of determining whether a trademark is distinctive.

Figure A7: Model Confidence and Publication Delay



Note. Publication delay versus predicted probability (model confidence) shows a negative relationship. The blue dots represent the average publication delay value in each decile.

A.4 Additional Outcome Labels

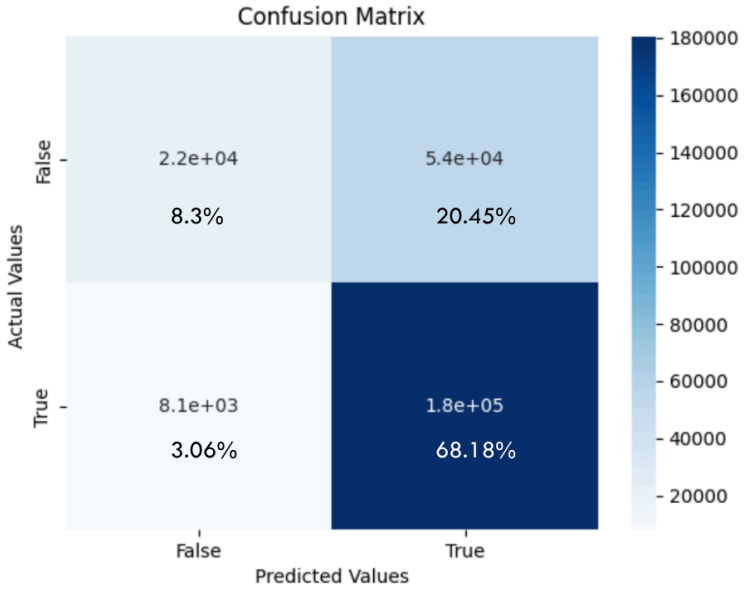
Other than using distinctiveness indicator as an outcome label, we created a publication indicator, a registration indicator and a used acquired distinctiveness indicator as additional outcome labels. Further, we created variants of these indicators (two for the registration indicator and one for the acquired distinctiveness indicator). We used the same methodology as described in Section 4(using DistilBERT to predict these additional outcome labels). Here, we describe the results from additional outcome labels and their variants.

A.4.1 Publication Indicator

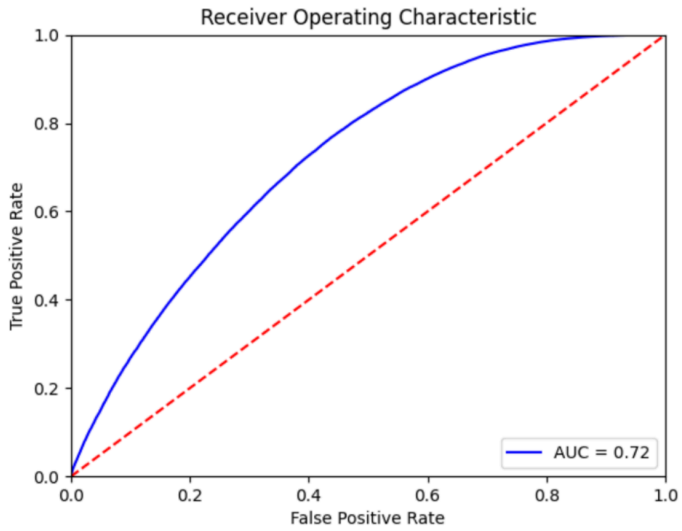
According to the USPTO, after examining an application, the trademark is approved for publication in the weekly online USPTO Trademark Official Gazette. We create the publication indicator on our dataset by examining whether the mark has a publication date. In our dataset, approximately 75% of the marks were published (having a label of 1). As done for

distinctiveness indicator, Figure A8 represents additional performance metrics of fine-tuned DistilBERT on the publication indicator. Figure A9 shows a graphic representation of precision on the publication indicator on varying thresholds. Table A8 depicts the same in a tabular format.

Figure A8: Model Performance Metrics



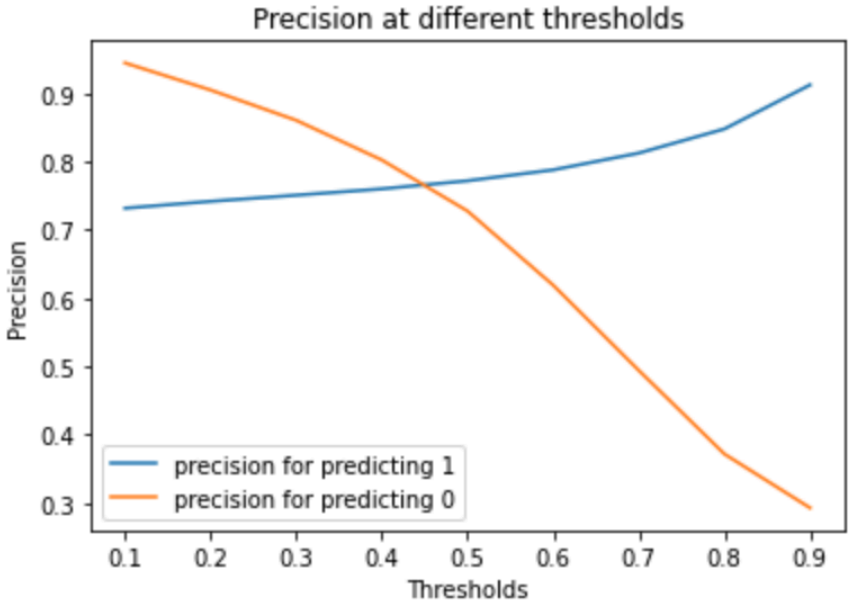
(a)



(b)

Note. Panel (a) shows the confusion matrix with a decision threshold of 0.5, and (b) shows the AUC of 0.72 achieved by the fine-tuned DistilBERT model on publication indicator

Figure A9: Model Precision at Different Thresholds



Note. A graphical representation of precision for publication (predicting 1) and no publication (predicting 0) of marks for varying decision thresholds (fine-tuned DistilBERT).

Table A6: Classification Metrics for Model Variants – Publication Outcome

Metrics	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Accuracy	0.77	0.77	0.77	0.77	0.77	0.77	0.77
Recall (weighted)	0.77	0.77	0.77	0.77	0.77	0.77	0.77
Precision (weighted)	0.76	0.76	0.76	0.77	0.76	0.76	0.76
F1 (weighted)	.73	.73	.73	.73	.73	.74	0.73
AUC	.72	.72	.72	.72	.72	.72	0.72
Mark	X	X	X	X	X	X	X
Statement	X	X	X	X	X	X	X
Translated Mark		X	X	X	X	X	X
Word in Dictionary			X	X	X	X	X
Mark Length				X	X	X	X
NICE Class					X	X	X
Pseudo Mark						X	X
NICE Description							X

Note. Fine-tuned DistilBERT classifier evaluation metrics for the publication outcome using a decision threshold of 0.5. We use weighted metrics to take into account the uneven distribution of mark publication in the dataset (75% of marks are published, and 25% not published)

A.4.2 Registration Indicator

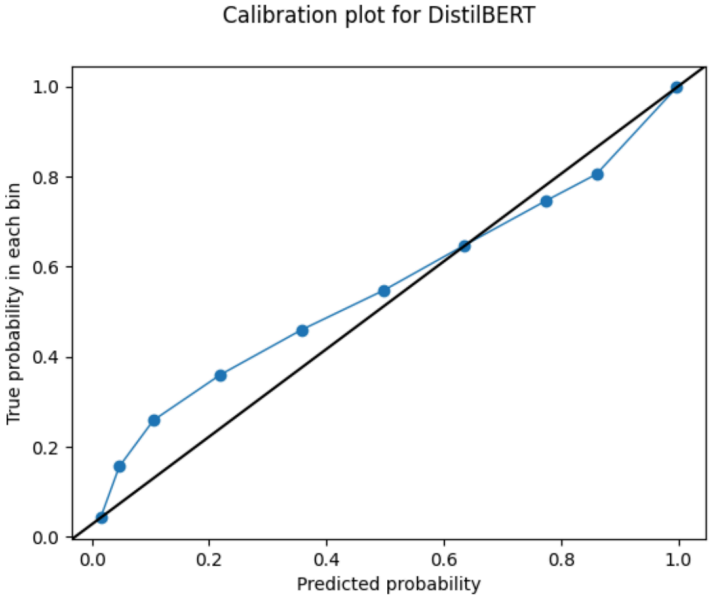
As described previously, publication of a trademark does not necessarily translate to being registered. We create a registration indicator on our dataset by examining if the mark has a registration date. In our dataset, approximately 58% of the marks had a registration date, and correspondingly a label of 1. Table A7 shows the classification metrics for the registration indicator by fine-tuning the DistilBERT model. Table A8 shows the precision for our model on different thresholds for decision making. Figure A10 shows the calibration plot for the model.

Users can still make a decision by referring to model precision at different thresholds (Table A8).

Table A7: Classification Metrics for Registration Indicator

Metrics	Value
Accuracy	0.63
Recall (weighted)	0.62
Precision (weighted)	0.65
F1 (weighted)	0.61
AUC	0.69

Figure A10: Calibration Plot for Fine-tuned DistilBERT



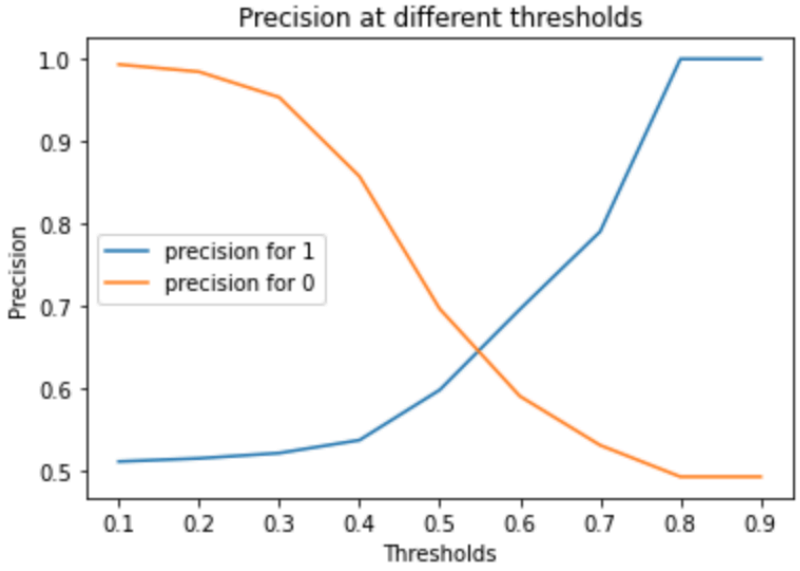
Note: Calibration plot with the true rate, binned by predicted probability..

Table A8: Precision at Different Threshold Levels

Decision Threshold	% data < threshold	% data > threshold	Precision 0 < threshold	Precision 1 > threshold
0.1	0.8%	99.2%	98.50%	51.60%
0.2	1.5%	98.5%	96.99%	51.93%
0.3	3.1%	96.9%	93.04%	52.63%
0.4	8.3%	91.7%	83.72%	54.36%
0.5	29.3%	70.7%	69.72%	59.89%
0.6	63.0%	37.0%	59.28%	69.02%
0.7	85.6%	14.4%	53.27%	77.81%
0.8	99.6%	0.4%	48.97%	90.65%
0.9	99.9%	0.1%	48.86%	99.74%

Note. Precision at different threshold levels for fine-tuned BERT Model on registration indicator after calibrating model output probabilities.

Figure A11: Graphical Representation of Table A8



Note. A graphical representation of precision for registration (predicting 1) and no registration (predicting 0) of marks for varying decision thresholds.

A.4.3 Registration Indicator: Filtering Intent to Use Applications

As a variant, we try to predict the registration indicator by dropping data points having an intent to use (current or at the time of filing). This field is already present in the USPTO Case Files dataset. We fine-tune the DistilBERT model on this dataset. We see that by dropping these rows, the accuracy, weighted precision and weighted recall for the model increase. Table A9 shows the classification metric for this model. Figure A12 shows the calibration plot, and Figure A13 shows the precision at different thresholds. Similar to the registration indicator, we see that the probabilities are inflated for lower values, and deflated for higher values.

Table A9: Classification Metrics for Registration Indicator, Filtering Out ITU Applications

Metrics	Value
Accuracy	0.74
Recall (weighted)	0.74
Precision (weighted)	0.74
F1 (weighted)	0.69
AUC	0.69

Figure A12: Calibration Plot for Tine-tuned BERT on Registration Indicator, Filtering Out ITU Applications

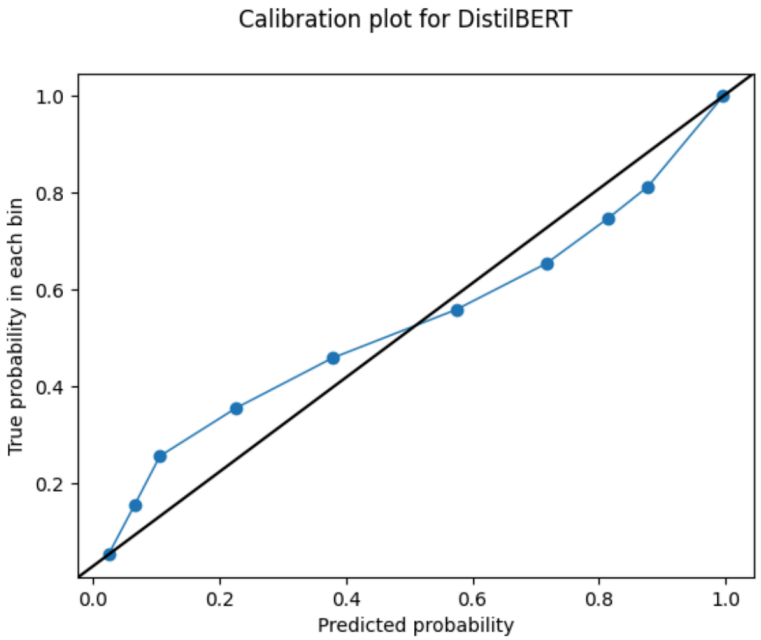
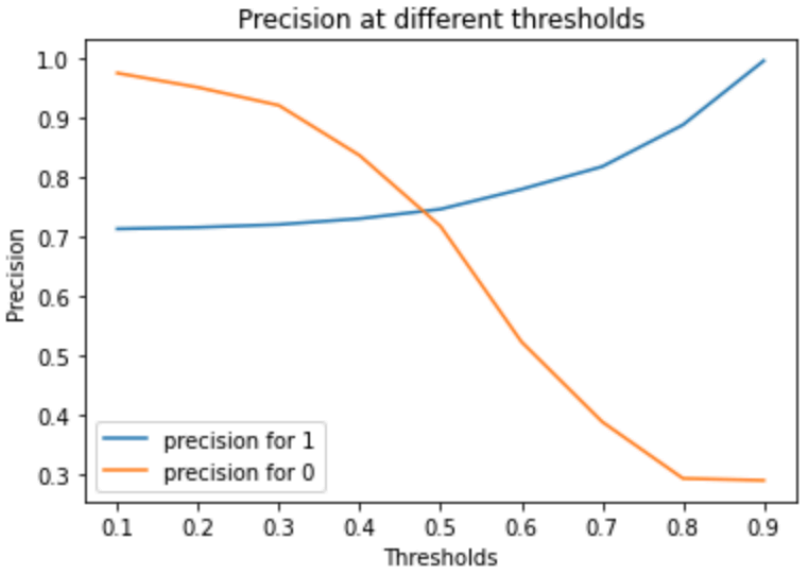


Figure A13: Precision at different thresholds: DistilBERT, Registration without ITU



Note. A graphical representation of precision for registration without ITU (predicting 1) and no registration without ITU (predicting 0) of marks for varying decision thresholds

A.4.4 Registration Indicator: Filtering Acquired Distinctiveness

As another variant, we filter out marks having an acquired distinctiveness, and again fine-tune DistilBERT model. Table A10 shows the classification metric for this model. Figure A14 shows the calibration plot, and Figure A15 shows the precision at different thresholds. Similar to the registration indicator, we see that the probabilities are inflated for lower values, and deflated for higher values.

Table A10: Classification Metrics for Registration Indicator, Filtering Out Acquired Distinctiveness Applications

Metrics	Value
Accuracy	0.65
Recall (weighted)	0.63
Precision (weighted)	0.65
F1 (weighted)	0.61
AUC	0.70

Figure A14: Calibration Plot for BERT, Filtering Out Acquired Distinctiveness Applications

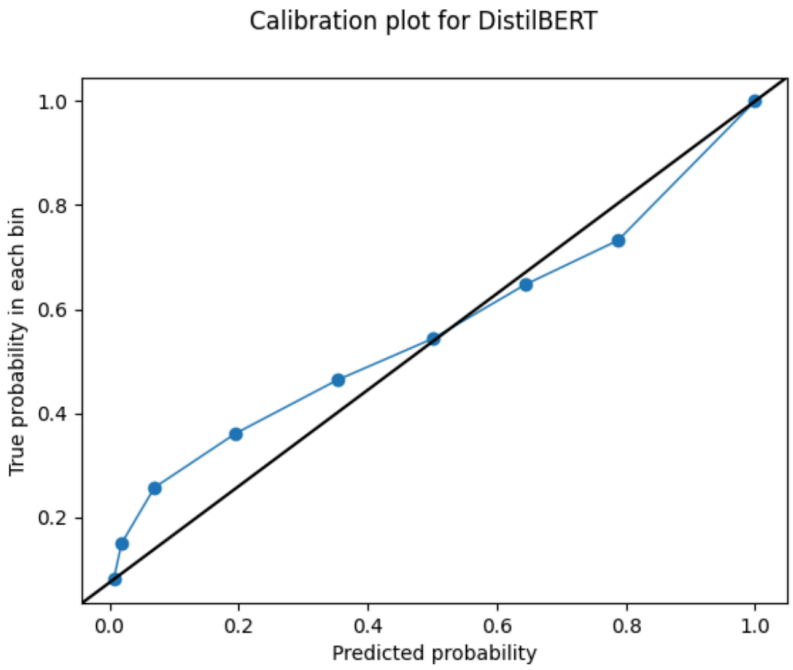
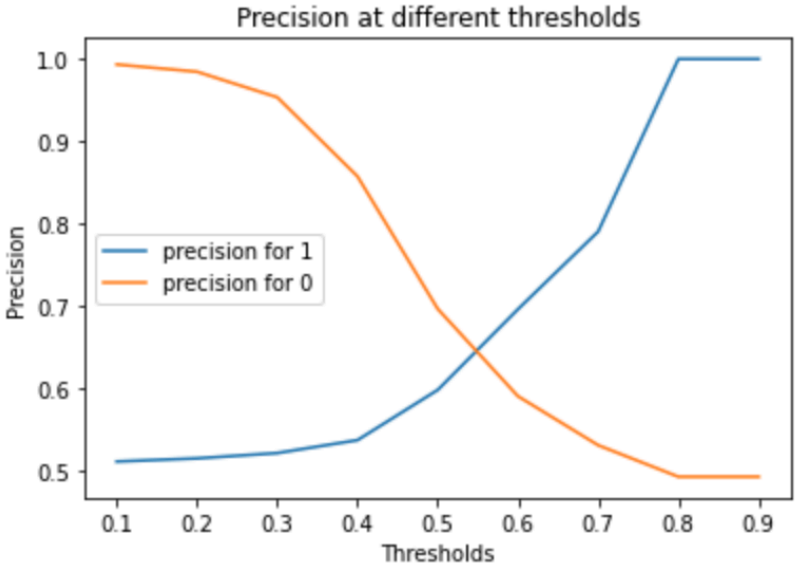


Figure A15: Graphical Representation of Precision for Registration Without Acquired Distinctiveness (predicting 1) and No Registration Without Acquired Distinctiveness (Predicting 0) of Marks for Varying Decision Thresholds



A.4.5 Acquired Distinctiveness Indicator

Similar to previous outcome labels, we try to predict acquired distinctiveness indicator (already available in USPTO Case files) using the same methodology. However, due to a very low incidence rate (only 1.80% applications had acquired distinctiveness), the results were not satisfactory. Table A11 shows the classification metrics for the fine-tuned BERT model. Here, instead of measuring weighted recall and precision, we focus only on the positive labels due to the low incidence rate.

Table A11: Classification Metrics for Acquired Distinctiveness Indicator

Metrics	Value
Accuracy	0.78
Recall (only 1s)	0.67
Precision (only 1s)	0.04
F1 (only 1s)	0.07
AUC	0.80

Note. Although the model has a decent recall, precision and F1 scores are pretty low.

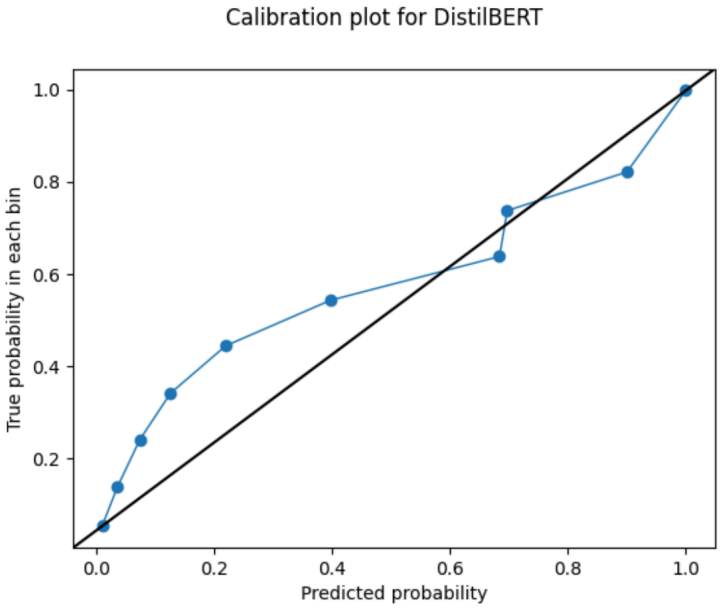
A.4.6 Acquired Distinctiveness Indicator: Filtering Registered Applications

As a variant outcome label, we filter out registered applications (registration indicator = 1), but not having acquired distinctiveness (acquired distinctiveness indicator = 0), while predicting acquired distinctiveness. This increases the incidence rate in our dataset from 1.80% to 4.22%. Although 4.22% is a low incidence rate, we again fine-tune DistilBERT on this dataset. We see a sharp increase in accuracy and precision, although recall falls down. Table A12 shows the classification metrics for the model. Figure A16 shows the calibration plot for the fine-tuned BERT model.

Table A12: Classification Metrics for the Acquired Distinctiveness Indicator, Filtering Out Registered Applications, But Not Having Acquired Distinctiveness

Metrics	Value
Accuracy	0.97
Recall (only 1s)	0.11
Precision (only 1s)	0.49
F1 (only 1s)	0.11
AUC	0.80

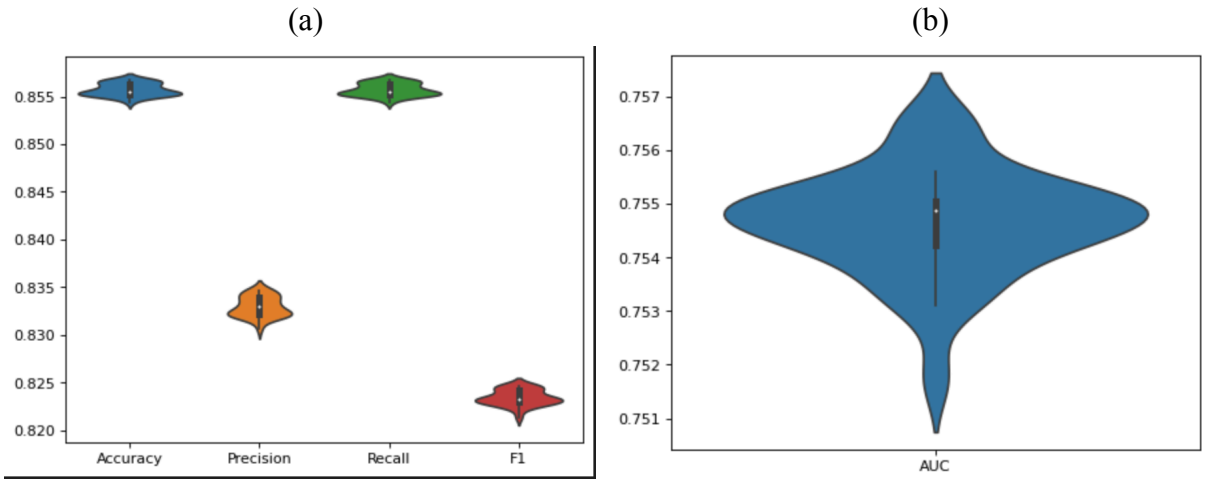
Figure A16: Calibration Plot for DistilBERT: Acquired Distinctiveness Indicator After Filtering Out Registered Applications, But Not Having Acquired Distinctiveness



A.5 Bootstrapping Performance Metrics

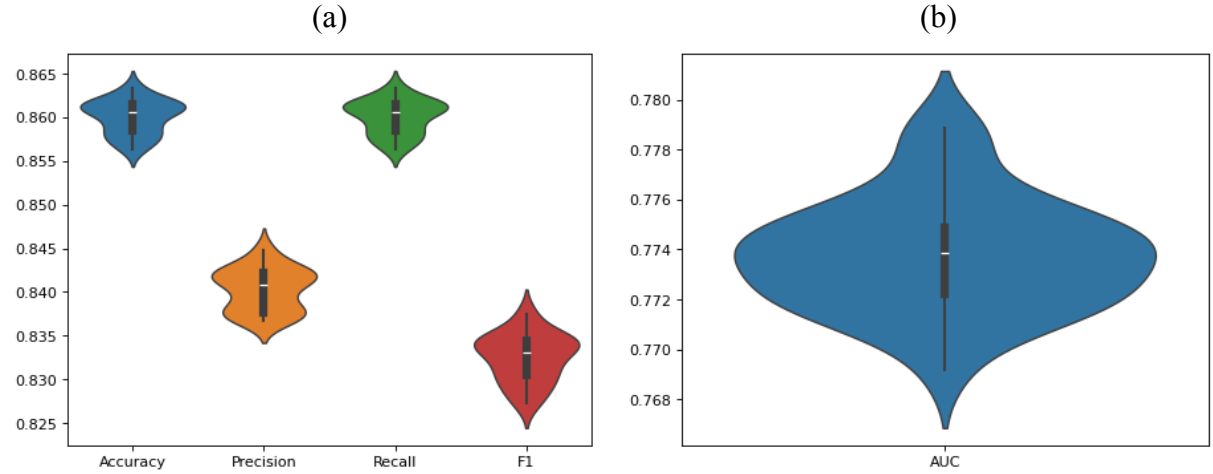
To assure model robustness, we perform bootstrapping and calculate confidence intervals of model metrics across 30 rounds with replacement on the test dataset. As seen from Figures A17 to A19, all of the metrics have low standard deviations.

Figure A17: Bootstrap performance of DistilBERT.



Note. X-axis represents the metric and Y-axis represents the score. (a) shows bootstrapping for Accuracy, Precision, Recall and F1 Score. (b) shows bootstrapping for AUC.

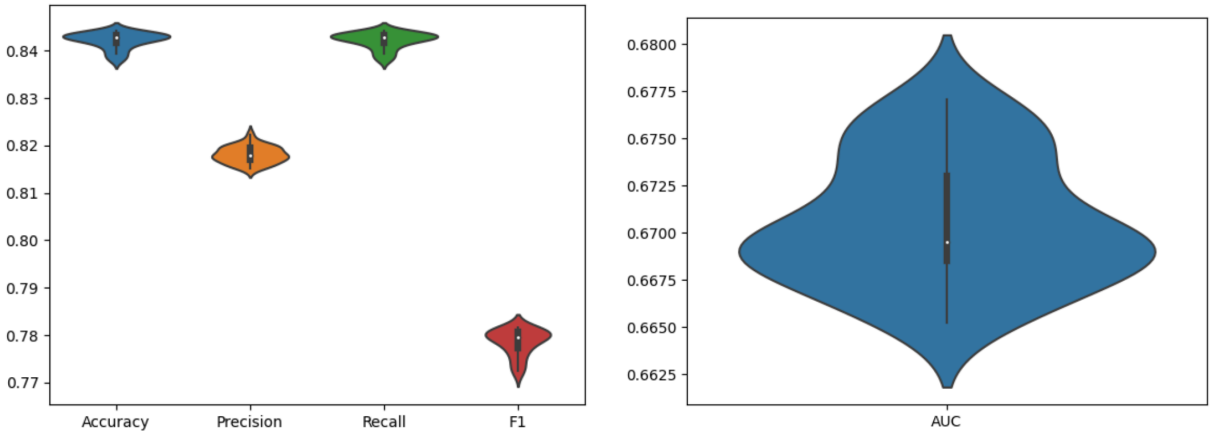
Figure A18: Bootstrap performance of RoBERTa



Note. The x-axis represents the metric and Y-axis represents the score. (a) shows bootstrapping for Accuracy, Precision, Recall and F1 Score; (b) shows bootstrapping for AUC.

Figure A19: Bootstrap performance of Xgboost

(a) (b)



Note. The x-axis represents the metric and Y-axis represents the score. (a) shows bootstrapping for Accuracy, Precision, Recall and F1 Score; (b) shows bootstrapping for AUC.

A.6 Further Robustness Checks

We perform three model checks for analyzing the performance with varying inputs. Firstly, we randomly shuffle the input order to check if there is an increase in performance based on the order of input. Secondly, we analyze how powerful the model is when using only the mark name to predict distinctiveness indicator and thirdly, we check the model performance when using all the features except the mark name to analyze the importance of other features. Table A13 summarizes the model performance in these three cases. The performance of randomly shuffled input is similar to the model performance in Table 3, suggesting no effect of shuffling the input order. The model performance using only mark, and all features except mark are worse than the performance in Table 3, suggesting both input types are required for a better performance. Bootstrapping performance metrics did not show any significant variation.

Table A13: Performance Summary for Different Input Types

Input Type	Accuracy	AUC
Randomly Shuffled	0.85	0.76
Only Mark	0.85	0.69
All features except Mark	0.84	0.71

A.7 Xgboost and fastText

We choose Xgboost to create a baseline model, and compare the results with our fine-tuned DistilBERT. Xgboost needs the inputs to be in numeric format. Inspired by Hemel & Ouellette (2021), we try to quantify the distance between a mark description and mark text, and further use it to predict our outcome labels. We use fastText embeddings to create a 300 dimension numeric representation of mark text and its description. fastText creates sub-word embeddings and is able to handle out of vocabulary words, making it a good choice to generate word vectors for trademarks, which are often made up words. Next, we take the difference between these two representations. If the embedding vectors of mark description and mark text are far away, we can assume that mark is more probable to get published, registered, or acquire distinctiveness. Further, we also use NICE classes, wordnet indicator and mark length as features to predict the outcome labels. Although fine-tuned DistilBERT outperforms Xgboost for all the labels, we do see decent results from this method as well. Table A14 shows a consolidated report on the classification metrics for different outcome labels.

Table A14: Performance of Xgboost on Different Outcome Labels

Model	Accuracy	Precision	Recall	F1	AUC
Publication Ind.	0.72	0.72	0.72	0.62	0.63
Registration Ind.	0.55	0.61	0.55	0.45	0.60
Registration Ind. (ITU Filtered)	0.71	0.71	0.72	0.60	0.60
Registration Ind. (Acq. Dist. 1 Filtered)	0.55	0.55	0.61	0.45	0.62
Acq. Distinctiveness Ind.	0.76	0.67	0.07	0.04	0.78
Acq. Distinctiveness Ind. (Filtered Registration=1)	0.97	0.03	0.46	0.05	0.78

A.8 Performance on Out-of-Vocabulary Words

As an experiment, we also analyze the performance for fine-tuned DistilBERT for predicting distinctiveness of a mark when the mark is not in our reference (Wordnet) dictionary. Table A15 shows the corresponding results. Compared to Table 3, we see an increase in accuracy and recall, but a decrease in AUC.

Table A15: Performance of fine-tuned DistilBERT on Out-of Vocabulary Words at a Decision Threshold of 0.5

Metrics	Value
Accuracy	0.89
Recall (weighted)	0.89
Precision (weighted)	0.86
F1 (weighted)	0.85
AUC	0.71

A.9 Adding Similar Marks Using fastText

Using fastText vector representation of a mark (mark text), we try to calculate similar marks and add it as an input feature. For every mark, similar marks were calculated within the same NICE Class using cosine distance between word vectors as a similarity metric. However, we did not see any improvement in the results.

A.10 Additional Material on Model Explanation

In this appendix, we present further studies on analyzing model explanations. In section A.10.1, we analyze the SHAP scores for samples having different levels of probabilities of being distinctive. In section A.10.2, we look at another model explanation method, BERTViz, focused towards the attention mechanism for transformer architectures, such as DistilBERT.

A.10.1 SHAP on Varying Levels of Model Probabilities

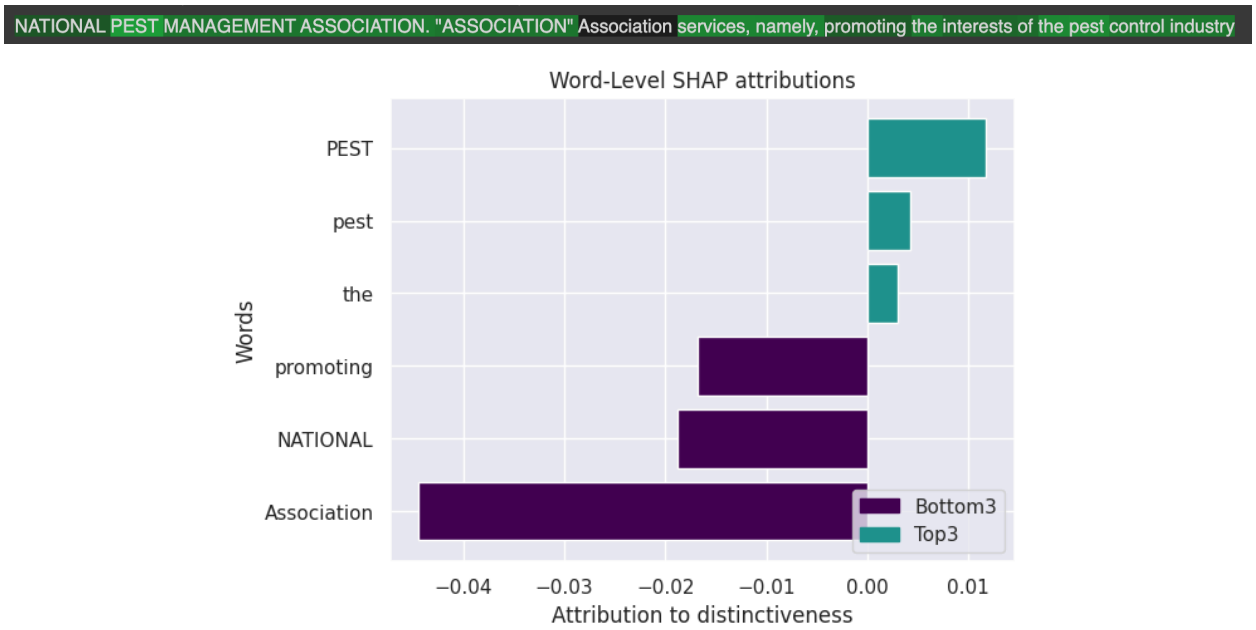
In Section 4.3, we demonstrated how SHAP scores can be used to explain which features of a trademark application contributed to our model's prediction. Going beyond the ZIPSCENE example used in that section, we now want to enhance our understanding of model explanations through SHAP by analyzing SHAP scores for marks with varying model probabilities. We begin by dividing the predicted probabilities into ten equal groups, or deciles, and select a sample of trademark applications from each decile for further SHAP analysis. These samples are intended to represent the overall data distribution. Figures A20 (a) to (j) show the SHAP scores for actual trademarks across varying levels of model probabilities, each associated with increasing levels of deciles. We notice that for probabilities less than 0.5, the overall mark contribution is often negative, where the individual words in the mark name have opposite directions of attributions. As we move towards higher probabilities, the individual words in the mark name are more aligned, contributing positively towards the mark being distinctive. Overall, this indicates that as

the predicted probabilities increase, the influence of the mark names on these probabilities shifts from negative to positive.

Figure A20

(a)

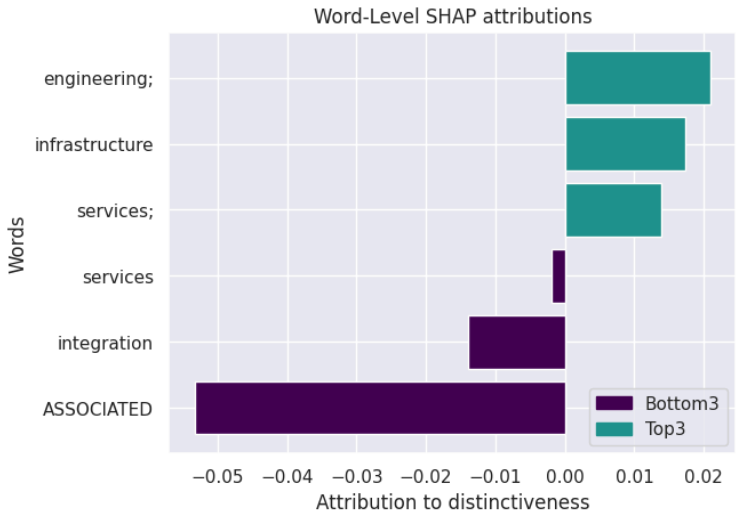
NATIONAL PEST MANAGEMENT ASSOCIATION ; Decile 1 ; Prob(distinctiveness) = 0.003



Note. Although the word "PEST" contributes positively towards distinctiveness, the word "NATIONAL" contributes negatively with a higher score, making the overall effect of the mark name negative.

(b)

ASSOCIATED CONSULTANTS ENTERPRISE SOLUTIONS ; Decile 2; Prob(distinctiveness) = 0.10

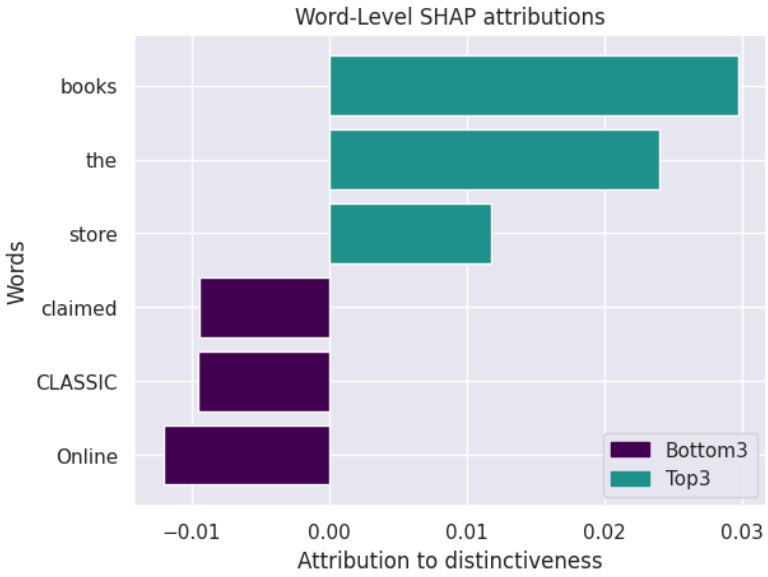


Note. The word "ASSOCIATED" hsa the highest negative contribution and none of the other words present in the mark name has a high positive contribution, making the overall effect of the mark name negative.

(c)

CLASSIC EDITIONS ; Decile 3 ; Prob(distinctiveness) = 0.25

CLASSIC EDITIONS. Color is not claimed as a feature of the mark. The mark consists of the word CLASSIC centered above the word EDITIONS. Online retail store services featuring books

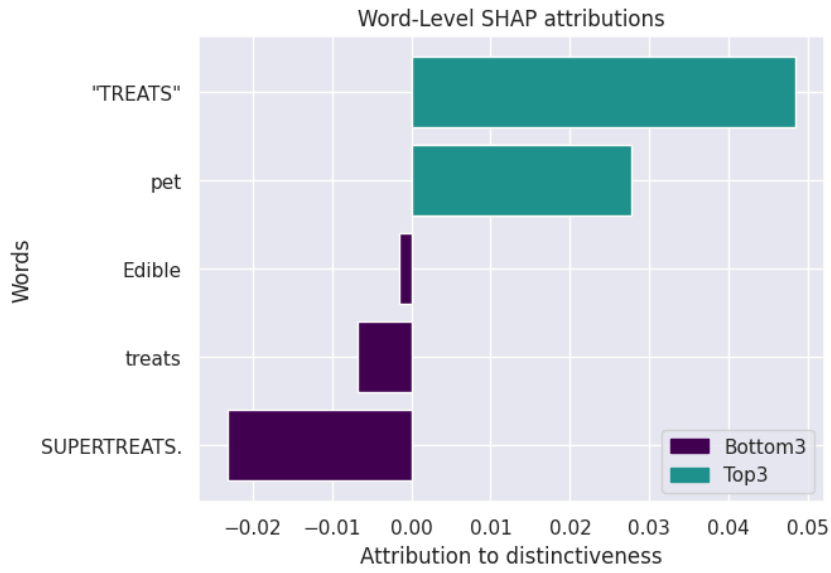


Note. The word "CLASSIC" has a high negative contribution and none of the other words present in the mark name has a high positive contribution, making the overall effect of the mark name negative.

(d)

SUPERTREATS ; Decile 4; Prob(distinctiveness) = 0.38

SUPERTREATS. "TREATS" Consumable pet chews; Edible pet treats

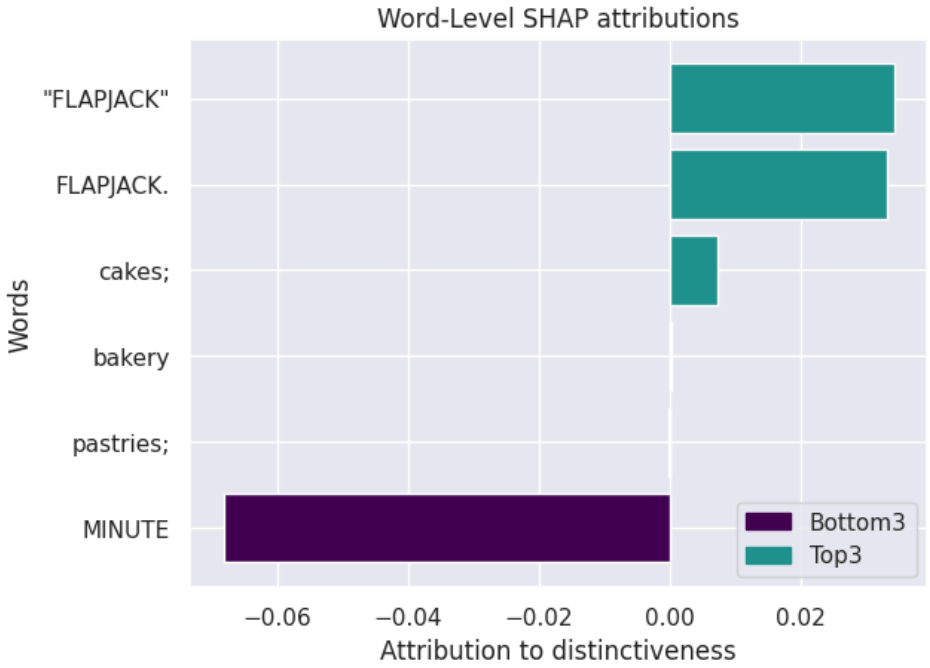


Note. The word "SUPERTREATS" has a high negative contribution. Besides, only two words - "TREATS" and "pet" contribute positively towards the mark being distinctive, resulting in a low probability of distinctiveness.

(e)

MINUTE FLAPJACK ; Decile 5; Prob(distinctiveness) = 0.48

MINUTE FLAPJACK. "FLAPJACK" Mixes for bakery goods; Pancake mixes; cake mixes; cupcakes mixes; mixes for grain-based pastries; cakes; pastries; Snack cakes; Grain-based snack foods

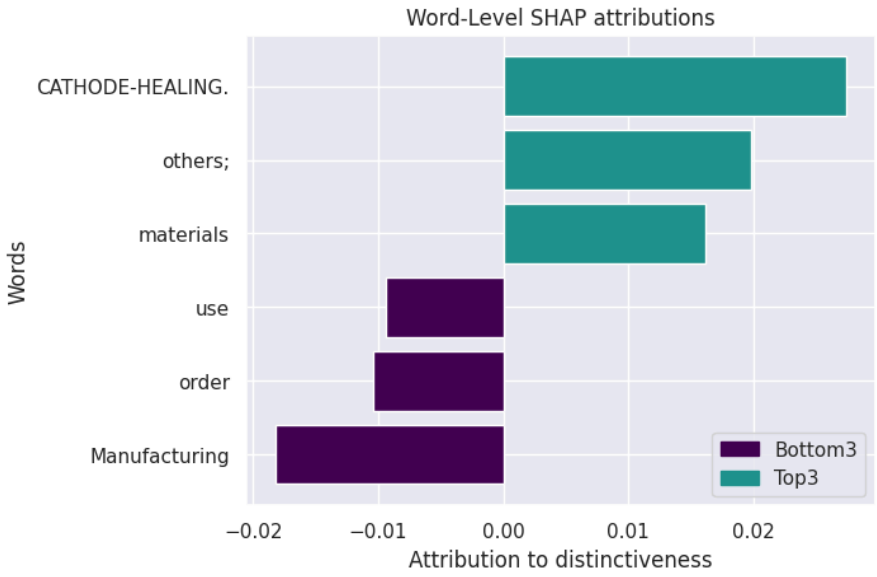


Note. The word "MINUTE" has a high negative contribution even though "FLAPJACK" has a positive contribution towards distinctiveness. Essentially, "MINUTE" outweighs "FLAPJACK", resulting in a distinctiveness probability of 0.48

(f)

CATHODE-HEALING ; Decile 6; Prob(distinctiveness) = 0.51

CATHODE-HEALING. Manufacturing of materials for use as battery cathodes to the order and specification of others; recycling services for battery cathode materials

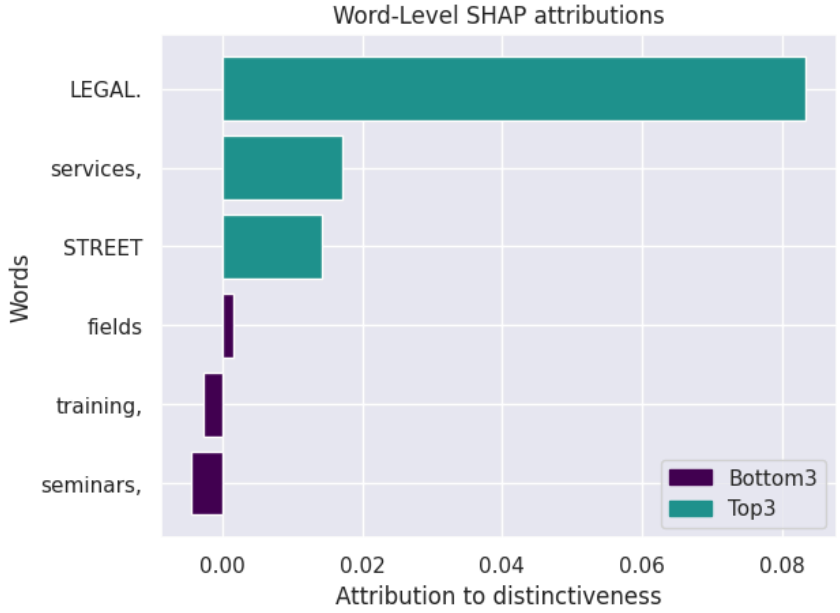


Note. The entire mark name "CATHODE-HEALING" has a high positive contribution.

(g)

STREET LEGAL ; Decile 7 ; Prob(distinctiveness) = 0.64

STREET LEGAL. Educational services, namely, providing live and on-line courses, workshops, lectures, seminars, training, and presentations in the fields of law enforcement, public safety, and other government agencies

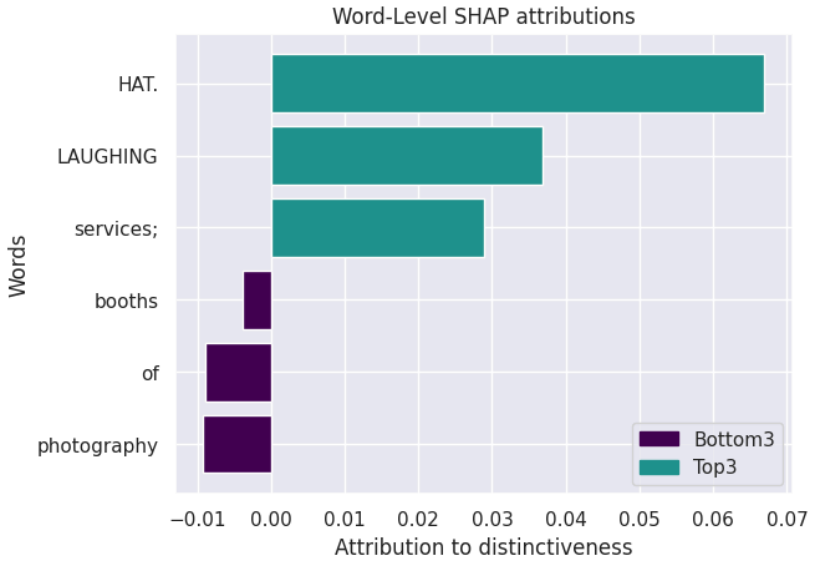


Note: Both the words "STREET" and "LEGAL" in the mark contribute positively towards the mark being distinctive.

(h)

LAUGHING HAT ; Decile 8 ; Prob(distinctiveness) = 0.77

LAUGHING HAT. Photography services; Rental of portable photography and/or videography booths for taking of pictures and videos

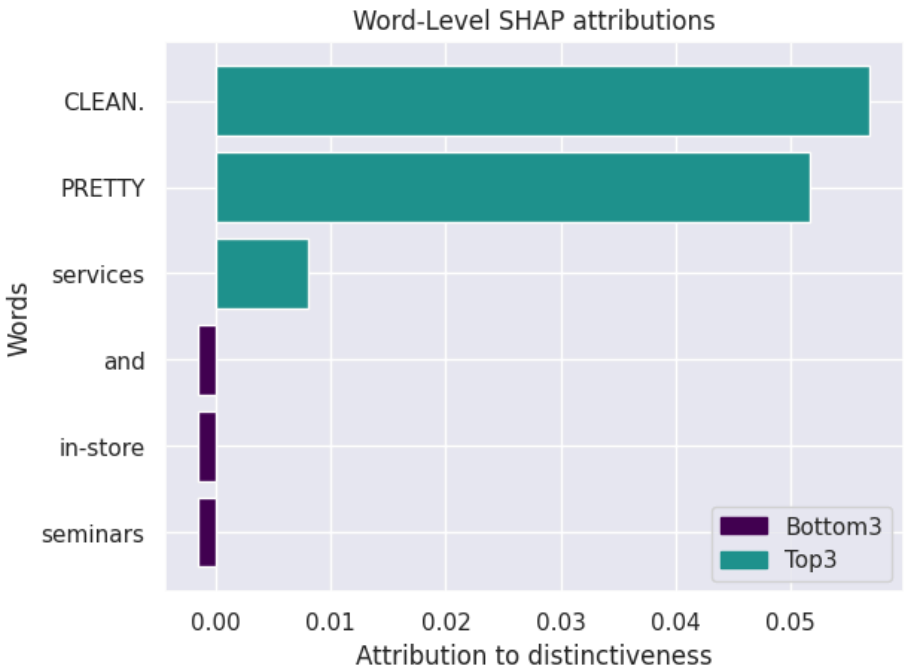


Note: Both the words “LAUGHING” and “HAT” in the mark name contribute positively towards the mark being distinctive.

(i)

PRETTY CLEAN ; Decile 9 ; Prob(distinctiveness) = 0.84

PRETTY CLEAN. Provision of training services in the field of beauty, cosmetics and skincare; Educational services, namely, online and in-store seminars in the field of beauty, cosmetics and skincare; information and advice relating to the foregoing services including online; Provision of training services in the field of beauty, cosmetics and skincare; Educational services, namely, providing online tutorial sessions in the field of cosmetics, skin care and beauty; Arranging of educational demonstrations; Audio-visual display presentation services for educational purposes, namely, providing non-downloadable webinars in the field of cosmetics, skin care and beauty; Providing online training services in the field of cosmetics, skin care and beauty; Audio-visual display presentation services for entertainment purposes, namely, provision of non-downloadable films and movies via a video-on-demand service; Entertainer services, namely, live visual and audio performances by a presenter in the field of cosmetics, skin care and beauty; Entertainment by means of live road shows in the field of cosmetics, skin care and beauty; Providing live educational demonstrations; Publication services, namely, publication of printed matter; Providing on-line non-downloadable publications in the nature of articles in the field of cosmetics, skin care and beauty; Publication of educational materials, namely, articles; Publication of instructional literature; Providing on-line non-downloadable electronic publications in the nature of articles in the field of cosmetics, skin care and beauty; Publishing of electronic publications; Publishing of educational and training guides; Publishing of printed matter in electronic form on the Internet

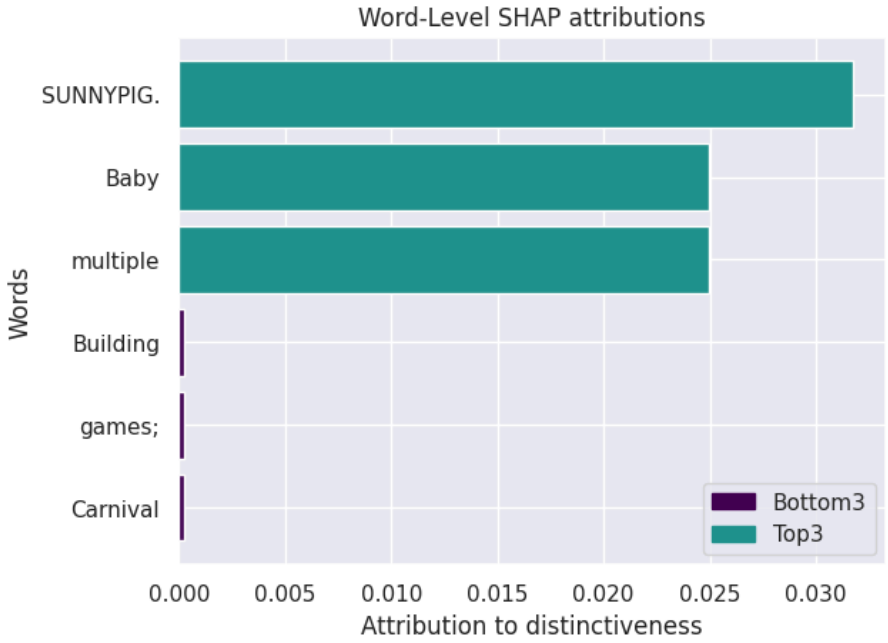


Note: Both the words "PRETTY" and "CLEAN" in the mark name contribute extremely positively towards the mark being distinctive.

(j)

SUNNYPIG ; Decile 10 ; Prob(distinctiveness) = 0.98

SUNNYPIG. Baby multiple activity toys; Bath toys; Building games; Carnival masks; Children's multiple activity toys; Controllers for toy prank toys; Controllers for toy cars, race cars, airplanes, boats; Crib mobiles; Doll houses; Doll house furnishings; Dolls' rooms; Dolls' beds; Dolls' houses; Dolls; Electronic novelty toys, namely, toys that electronically record, play back, and distort or manipulate voices and sounds; Jigsaw puzzles; Knee guards for athletic use; Modeled plastic toy figurines; Novelty toy items in the nature of prank toys; Radio-controlled toy vehicles; Radio controlled toy cars, race cars, airplanes, boats; Scale model kits; Scale model vehicles; Spinning tops; Stuffed toys; Teddy bears; Toy cameras; Toy houses; Toy models; Toy scooters; Toy vehicles; Toy building blocks; Toy cars; Toy figures; Toy masks; Toy telescopes; Toy watches; Whoopee cushions; Parlor games; Parlour games The wording "SUNNYPIG" has no meaning in a foreign language.



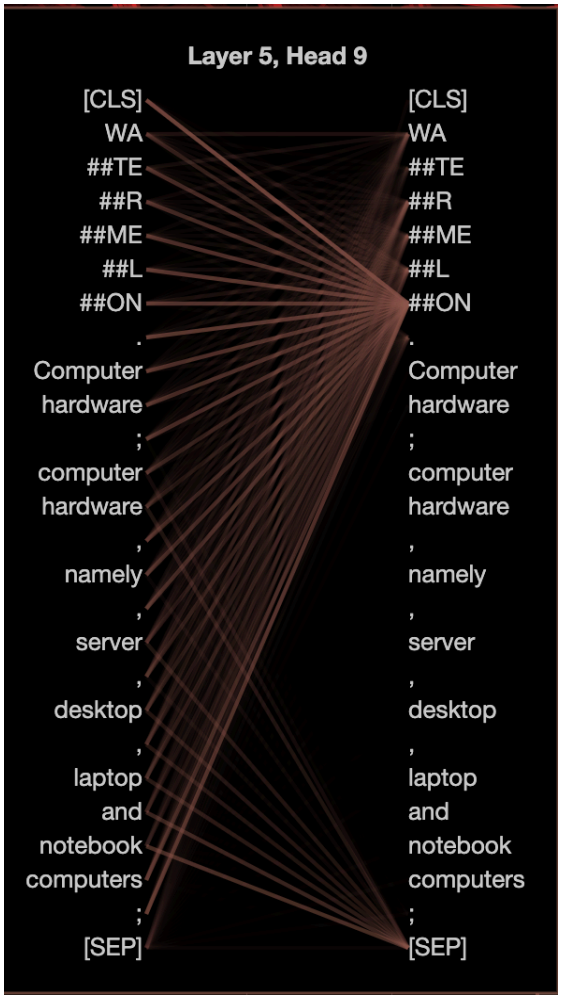
Note: The word "SUNNYPIG" contributes extremely positively towards the mark being distinctive.

A.10.2 BERTViz

SHAP is a model-agnostic method, which does not consider BERT's inner workings. In transformer-based models, attention scores provide the relevance of one word to another. Using an attention visualization tool such as BERTViz (Vig 2019) can help in understanding the model's behavior through attentions. Figure A21 visualizes the attention scores for the last layer (layer 5) and attention head 9 for "WATERMELON" in the context of computer hardware and fruit distributorship. Figure A22 shows the attention scores distributed for "computers" and "fruit." We chose attention head 9 as it represented a "broad attention" distribution (Clark et. al. 2019). From Figure A22, it becomes apparent that attention distribution is focused more towards the mark name in the context of computer hardware than fruit distributorship. Thus, the overall relevance of the mark name decreases when in context with the fruit distributorship. The comparison between the attention distribution of "computers" and "fruit" in Figure A22 also shows that "computers" attend to "WATERMELON" more than "fruit." As we saw previously that the mark name is one of the most important features, decreasing its relevance will contribute to decrease in the model output probability. To present this numerically, we calculate the overall attention towards "WATERMELON" in the last layer for the two cases, where the average attention score of "WATERMELON" decreases from 0.1290 to 0.1082. The analysis provided further in this section for the made-up mark "XADERMAC" in Figure A23 extends this example to fanciful marks.

Figure A21: BERTViz Attention Distribution for "Watermelon"

(a) Context of computer hardware

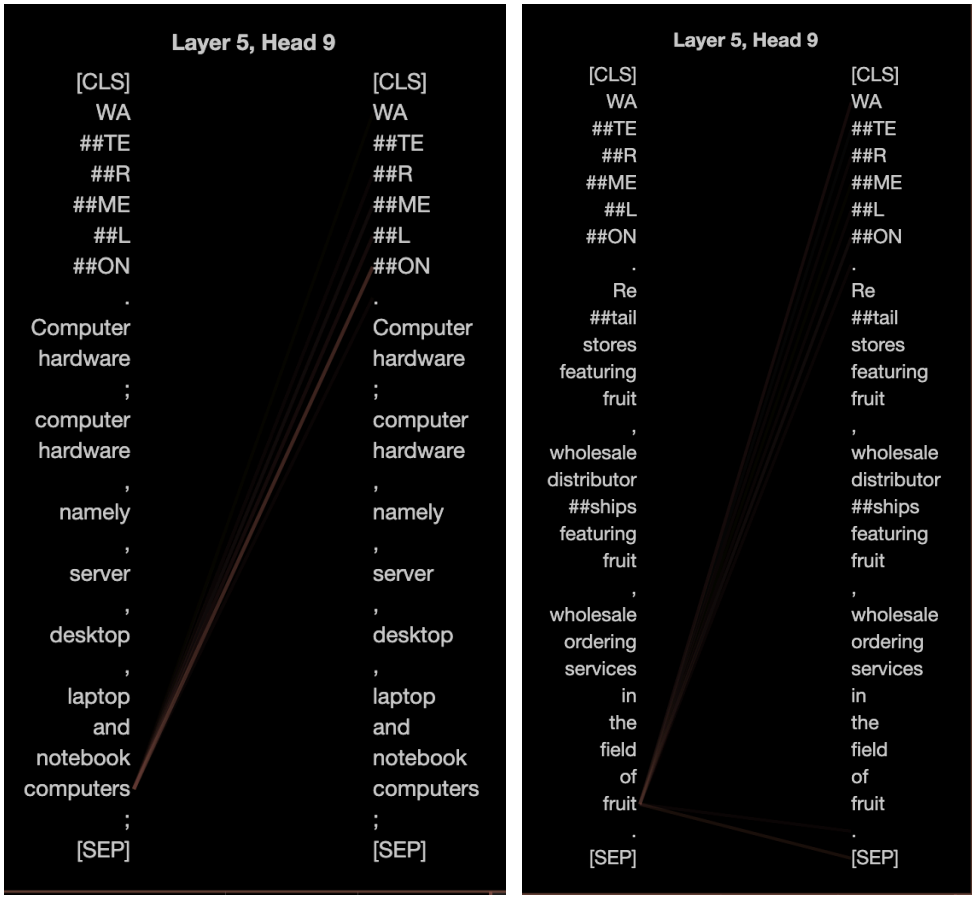


(b) Context of fruit distributorship

Note. We use BERTViz to visualize the attention distribution for "WATERMELON". (a) displays stronger attention towards mark name than (b). Darker lines implies stronger attention.

Figure A22: BERTViz Attention Distribution for "Watermelon"

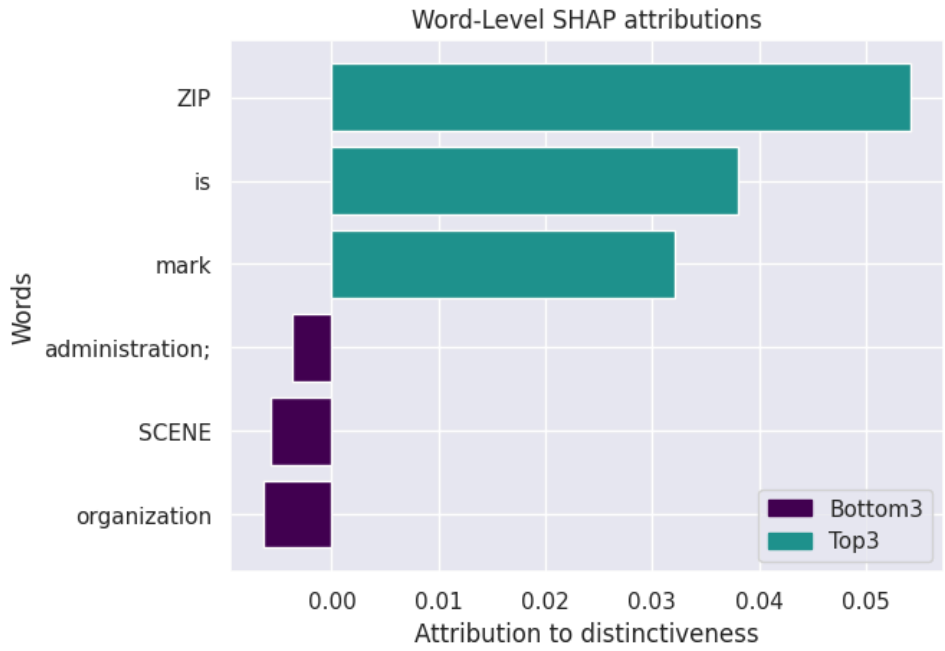
(a) Attention distribution for computers (b) Attention distribution for fruit



Note: BERTViz allows hovering over a word on the left to display the corresponding words it attends to through lines. (a) displays stronger attention of "computers" towards mark name than (b) "fruits" towards mark name. Darker lines implies stronger attention.

Figure A23: Shap Value Plots

(A) ZIPSCENE; P(distinctive) = 0.91



(B) XADERMAC; P(distinctive) = 0.80

XADERMAC. Retail stores featuring fruit, wholesale distributorships featuring fruit, wholesale ordering services in the field of fruit.

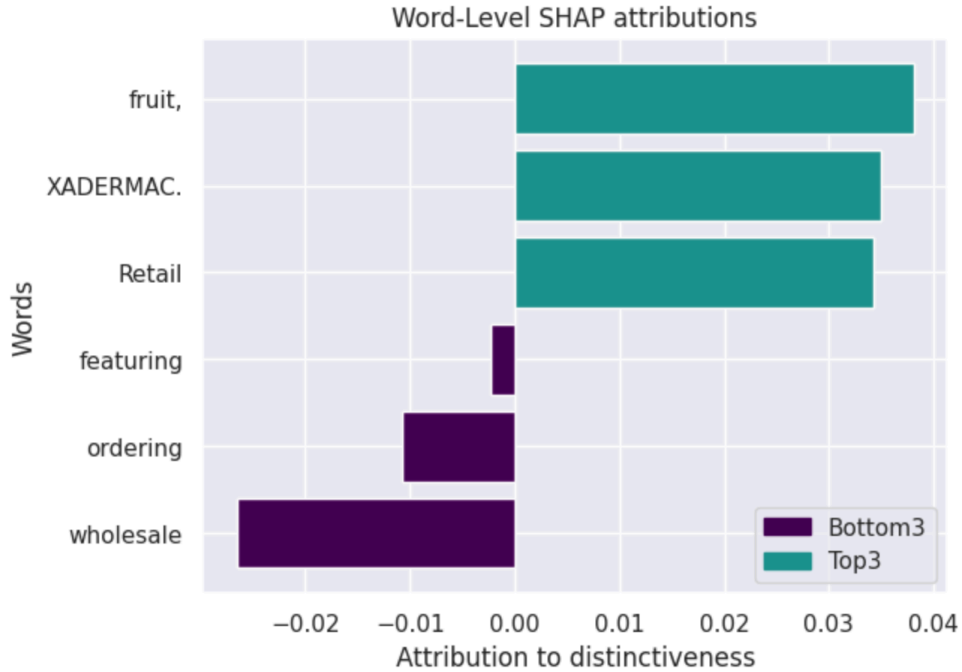
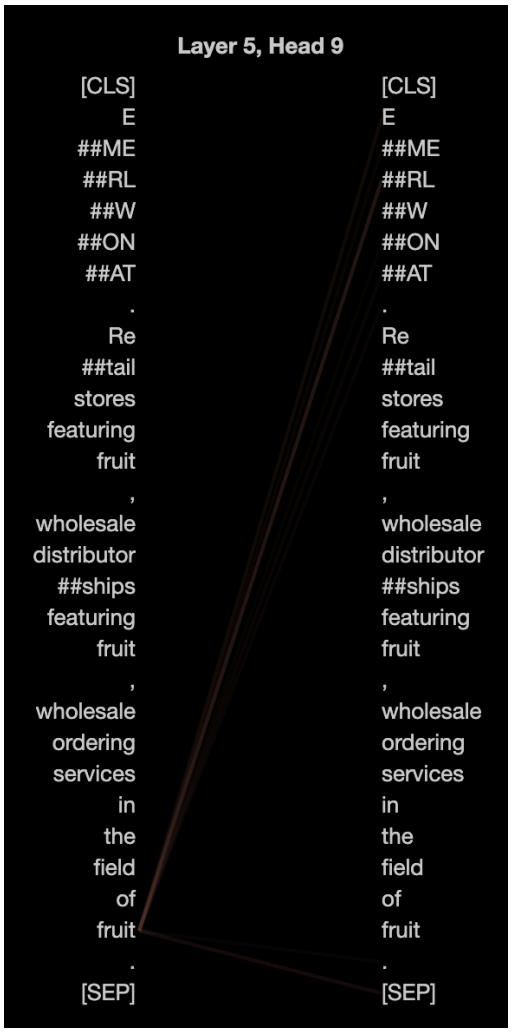


Figure A24: BERTViz Attention Distribution for EMERLWONAT

(a) "fruit" attending to mark name



(b) overall attention distribution

Note. Darker lines implies stronger attention.

Appendix B: Additional Figures and Tables

This section shows the distributions of quantifiable independent variables in our dataset, i.e, mark length (Figure B1) and NICE Class (Figure B2). Additionally, Table B1 shows the incidence rate (proportion of 1s) for all the indicators in our dataset.

Figure B1: Histogram of Mark Length

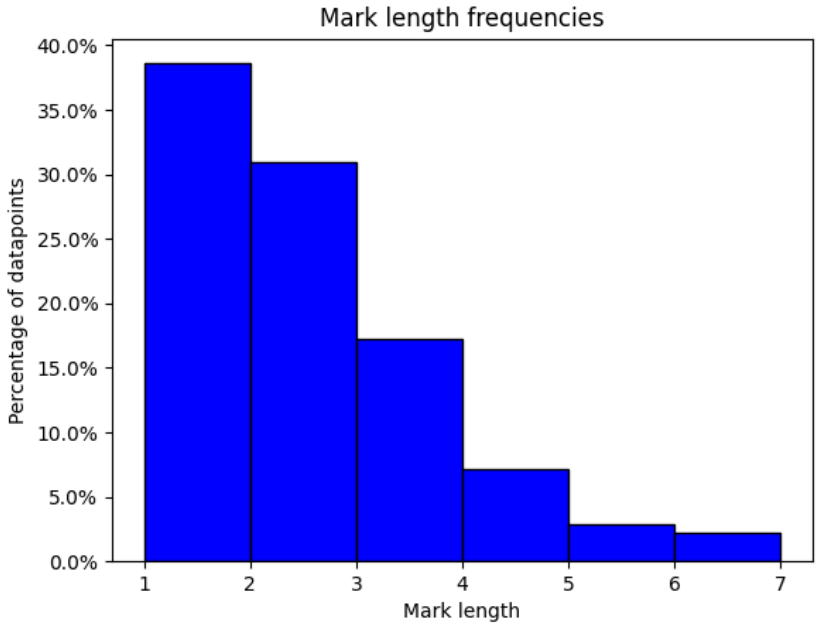


Figure B2: NICE Class Distribution

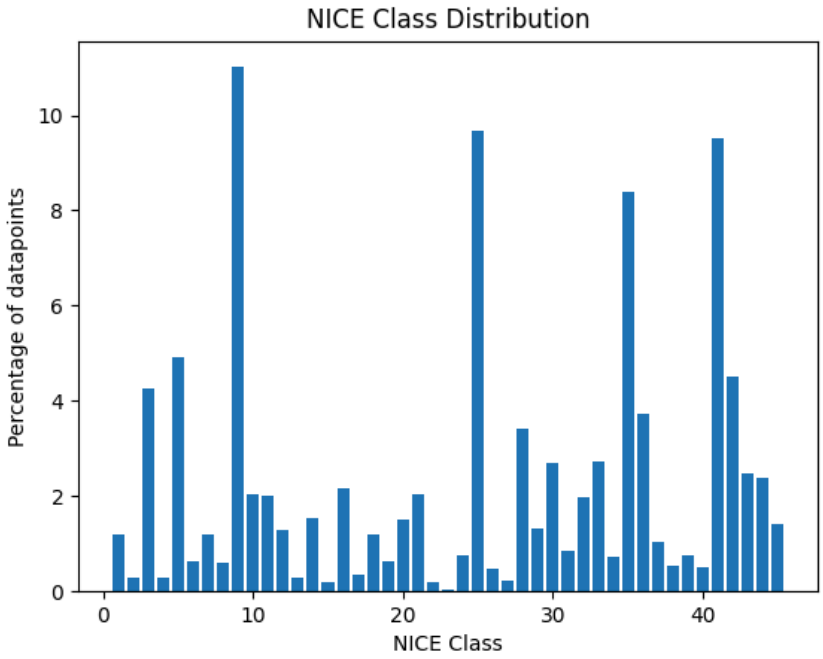


Table B1: Incidence Rate of All Outcome Labels

Outcome Label	Incidence
Acquired Distinctiveness	1.80%
Acquired Distinctiveness without Registration	4.22%
Registration Indicator	58.81%
Registration Indicator without Acquired Distinctiveness	58.38%
Registration Indicator with Intent to Use Applications filtered	76.72%
Publication Indicator	74.91%
Distinctiveness Indicator	84.01%