Understanding Brand Consistency from Web Content

Soumyadeep Roy
IIT Kharagpur, India
soumyadeep.roy9@iitkgp.ac.in

Niloy Ganguly
IIT Kharagpur, India
niloy@cse.iitkgp.ac.in

Shamik Sural
IIT Kharagpur, India
shamik@cse.iitkgp.ac.in

Niyati Chhaya
Adobe Research, India
nchhaya@adobe.com

Anandhavelu Natarajan
Adobe Research, India
anandvn@adobe.com

ABSTRACT

Brands produce content to engage with the audience continually and tend to maintain a set of human characteristics in their marketing campaigns. In this era of digital marketing, they need to create a lot of content to keep up the engagement with their audiences. However, such kind of content authoring at scale introduces challenges in maintaining consistency in a brand's messaging tone, which is very important from a brand's perspective to ensure a persistent impression for its customers and audiences. In this work, we quantify brand personality and formulate its linguistic features. We score text articles extracted from brand communications on five personality dimensions: sincerity, excitement, competence, ruggedness and sophistication, and show that a linear SVM model achieves a decent F1 score of 0.822. The linear SVM allows us to annotate a large set of data points free of any annotation error. We utilize this huge annotated dataset to characterize the notion of brand consistency, which is maintaining a company's targeted brand personality across time and over different content categories; we make certain interesting observations. As per our knowledge, this is the first study which investigates brand personality from the company's official websites, and that formulates and analyzes the notion of brand consistency on such a large scale.

KEYWORDS

brand personality; reputation management; affective computing; text classification

ACM Reference Format:
Soumyadeep Roy, Niloy Ganguly, Shamik Sural, Niyati Chhaya, and Anandhavelu Natarajan. 2019. Understanding Brand Consistency from Web Content. In *11th ACM Conference on Web Science (WebSci '19), June 30-July 3, 2019, Boston, MA, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3292522.3326048

1 INTRODUCTION

Organizations tend to maintain a personality or a set of human characteristics in their marketing campaigns, which help them to uniquely position themselves in a market segment and differentiate from other products. For example, Red Bull positions itself as courageous and outgoing and Nike portrays itself as athletic. Aaker [1] formalizes the brand persona dimensions highlighted in Table 1. Maehle [15] attempts to understand how consumers form their perceptions of the brand persona dimensions and also what product or brand characteristics influence these perceptions. There exist very few studies [12, 27] that attempt to quantify this implicit notion of brand personality. Brand personality is one of the dimensions that forms the brand image of an organization and it also significantly contributes towards the understanding of consumer choice [19].

Trait	Explanation
Sincerity	A brand which portrays itself as honest, friendly, sincere, or down–to–earth.
Excitement	A brand portraying itself as spirited, imaginative, trendy, and contemporary.
Competence	If a brand describes its successes and achievements in its articles it comes out as being competent. Competence in general is evoked when the reader interprets a brand's success from its content.
Sophistication	A brand which portrays itself as glamorous, charming, or catering to the upper class.
Ruggedness	A brand which portrays itself as adventurous, outdoorsy, tough, or Western.

Table 1: Brand Dimensions

In the era of digital marketing, brands need to create a lot of online content to keep up the engagement with their audiences. Brands also tend to share a lot of posts not directly promoting the brand such as information about the product domain and utility related insights to engage with its audience. This kind of online content authoring at scale introduces challenges in maintaining consistency in a brand's messaging tone which is very important from a brand's perspective to ensure a persistent impression on its customers and audiences. Monitoring and maintaining such brand consistency on a large scale is difficult, and require costly human experts.

To this end, the paper exploits several classification algorithms to check the brand persona of a content. Delin [6] describes how brand personality and brand value are built based on brand content. This research on content articles published by brands follows four linguistic frameworks: chains of reference, participant roles, presupposition and assumption, and tenor. We leverage these linguistic

WebSci '19, June 30-July 3, 2019, Boston, MA, USA

ACM ISBN 978-1-4503-6202-3/19/06…$15.00
https://doi.org/10.1145/3292522.3326048

features in our work to build a simple supervised classification model which needs annotated data. We scrapped the official websites of the Fortune 1000 companies of 2017 and accumulated around 30 gigabytes of textual data. We randomly select 600 articles and use Amazon Mechanical Turk for crowdsourcing the text annotation, thereby forming a human-annotated data set. With this gold standard dataset, a suite of standard classifiers are tested; out of which a linear SVM turns out to be the best achieving an F1-score of 0.822.

We classify the entire dataset (298112 web pages content covering 643 companies after several rounds of cleaning of the 30 GB data collected) using the selected linear SVM. Further, we consider a subset of the data which classify with a confidence value ≥ 0.095 and obtain a total of 93321 data points which covers web postings from 536 Fortune 1000 companies. The confidence threshold is considered to ensure the corresponding automatic annotations are not flawed. With the sanitized data MT_{large}, we now check the performance of companies with respect to their ability to maintain brand personae. As per our knowledge, this is the first study which attempts to quantify and investigate brand consistency in such a large scale. We conduct both temporal and non-temporal studies, find the type of brand personae displayed by different kinds of posts and identify the set of companies who are able to maintain their brand consistency over time.

The rest of the paper is structured as follows. In Section 2, we start with the prior art. Section 3 discusses the dataset collected and the process of cleaning the dataset as well as various observations regarding the dataset. Section 4 presents the proposed classification model followed by experimental results. In Section 5 we perform the large scale characterization study of brand consistency. The paper concludes with a summary of discussions and several directions for future work.

2 PRIOR RESEARCH

While sentiment and emotion analysis is extensively studied, unavailability of well-tagged datasets for other dimensions such as brand personality introduces a challenge in developing automated methods for detection and study of these dimensions.

Aaker [1] develops a theoretical framework for brand personality dimensions and lists five dimensions of brand personality. Xu et al. [27] present an approach for predicting perceived brand personality in social media, with the underlying hypothesis that brand perception depends on user imagery, employee imagery, and official announcements. Liu et al. [13, 14] further extend it to build a novel visual analysis tool, to help explain the association between brand personality and its above–mentioned driving factors on social media. Liu et al. [12] analyze how consumers and companies portray their brand perception in visual social media (images, instead of text) and attempt to capture the intangible differences between the two competing brands over the same product.

This issue of maintaining brand consistency and the cost associated with it is a well-established problem. This was a task in CLEF called RepLab [3], in which one focuses on monitoring the reputation of companies and individuals. However, the work is only on Twitter because it states that it is the critical media for early detection of potential reputational issues. The tweets are assigned to one of the seven standard reputation dimensions (performance, products & services, leadership, citizenship, governance, workplace, innovation) of the RepTrak Framework [1], which reflect the affective and cognitive perceptions of a company by different stakeholder groups [4]. Spina et al. [20] addresses this reputation monitoring problem faced by experts by formulating it as a topic detection task. Their work focuses only on Twitter.

Various features including LIWC [23], Mairesse [16], character-level features, and responsive patterns [22] have been found to be useful in predicting human personality from text. [27] finds that the income and needs of the consumer affect consumption behavior as well as their personality traits.

Muller and Chandon [17] studied the effect of a forced visit to a company website on brand personality. Since, companies strategically place verbal or non-verbal cues in the websites to evoke specific emotions from their users, Douglas et. al. [7] proposed the Website Emotional Features Assessment Model (WEFAM) based on website features like site activation, site affection, site confidence, site serenity, site superiority and site surgency.

Su et al. [21] train an RNN with LIWC and other grammatical features as input to predict personality trait scores. Wei et al. [25] use CNN with 1,2,3-grams kernels to capture structures in text. Yang et al. [28] build a hierarchical representation of documents, constructed from words to sentences and then to the document level. Ling et al. [10] goes deeper than word-level towards character-level, while Liu et al. [11] extends their work for short texts. However, due to the limited amount of annotated data, deep learning framework cannot be used in this context.

3 DATASET

We collected text content of the Fortune 1000 companies for the year of 2017 from their official websites. We only consider the following pages - *about the company*, *media releases*, *blogs and communication* directed towards the customers. We perform an extensive crawl using the Scrapy[2] framework and filter the pages based on keyword-based inclusion and exclusion rules over the webpage URL. We consider the web page types that aim to engage with the customer directly, first in terms of portraying their brand characteristics like about, history, vision, commitment, who-we-are and secondly, in terms of informative content like blogs, media releases, investors, newsroom. We limit our crawling task by not considering product pages like showroom, products, store, and content not targeted explicitly for consumers like legal, policy, disclaimer. The inclusion keywords are - about, about-us, news, press, introduction, strength, investors, history, vision, benefits, commitment, people, why-choose-us, who-we-are, approach, media, blog, social, while the exclusion keywords are - job, jcr_content, events, legal, help, showroom, products, store, project, career, policy, disclaimer, report. For each company, we start from their home page and limit ourselves to pages within the same domain name. We filter the websites that contain non-English text content. Given a web page, we parse only the ASCII text content with the paragraph ($< p > ...ASCIItext... < /p >$) HTML tags and concatenate

[1] $http : //www.reputationinstitute.com/about - reputation - institute/the - reptrak - framework$
[2] https://scrapy.org/

them together, separated by a paragraph separator marker. We are able to collect data from 299481 corporate web pages, covering 643 companies. We name this large dataset as MT_{large}.

3.1 Static and dynamic pages

Each of the company webpages consists of static and dynamic contents. Static pages are one which explicitly defines the brand which a company stands for like its mission, vision and core values. Generally, the frequency of such posts is quite less and mostly posted during the launch of the website. Dynamic web pages usually comprise of the content used for continually engaging with the audience. They are blogs, news, media or press releases, notes to investors posted at regular intervals. MT_{large} thus contains static pages and dynamic pages for 338 and 643 companies respectively. Static page content creates the brand impression while it is imperative that the dynamic content maintains that to ensure brand consistency. However, in this section, we perform some elementary data study to understand the nature of post of dynamic content.

Static keywords	introduction (34), about (573), commitment (45), people (252), vision (48), strength (429), history (1116), approach (571), benefits (930)
Dynamic keywords	media(19203), blog(36844), news(92448), press(52544), investors(5837)

Table 2: Keywords used to divide static and dynamic web content

3.2 Extracting temporal information from web content

We, therefore, need the timestamp information as to when a dynamic web page was posted by the company to understand the temporal behavior of the content. MT_{large} contain 298112 such dynamic posts covering 643 companies, from which we are able to timestamp information from only around 49.18%, accounting to 140,337 number of posts. We manually observe that the time-stamps usually have granularity in terms of days and weeks. 75.01% of these posts contain day-level information, while the remaining have a year as granularity. Here, we only consider company postings done between the period of January 2000 and September 2017.

3.3 Basic Observations

3.3.1 Volume of dynamic posts . We observe among the Fortune 1000 companies, the number of posts are roughly similar (Figure 1), although we find that there are occasional spikes representing companies who post way more than the average. These spikes are more prevalent among higher ranked companies.

3.3.2 Inter-arrival time between two postings. We analyze the inter-arrival time between two company postings and study the patterns that are prominent across different companies. By postings, we refer to only dynamic web content, and we consider the content where the date is present in the granularity of a day (52432 posts). We perform both sector-wise and industry-wide study. We observe in Figure 1 that the inter-arrival time posting patterns is heavy-tailed and the pattern is similar across sectors. We select the top 5 sectors with the most number of posts– technology (48219), financials (11739), energy (4915), healthcare (4685) and business services (3747), for depicting their heavy-tailed behavior. Here, we consider the dynamic posts whose granularity is in days and whose inter–arrival time is non-zero, In Figure 1, we observe peaks appearing consistently after an interval of 30 to 33 days and further investigate this issue, by studying what type of dynamic posts are more prevalent during these peaks. We observe that a significant value of 77.96% of them are posted at the end of the month. We consider day 1, 2, 30, 31 of a month as month-end. We observe the following proportion of post types - media (9.63%), blog (13.98%), news (66.95%), press (25.67%) and investor (24.0%) and the highest being the dynamic post type 'news' .

4 CLASSIFICATION

In order to check the brand persona of a content, we build up a classifier. To have a supervised model, we require annotated data as well as feature extraction from the dataset. These steps are discussed next followed by the classification models and experimental results.

4.1 Annotation

Each article is annotated in two different ways (a). annotators follow a 5 point Likert scale to annotate the articles on each of the five dimensions of brand personality (similar to the one used by [9] and [18] for annotating formality), (b). they rank the five dimensions in the order in which they are evoked from a given article. Each article is annotated 3 times by different annotators. Consistency between the two methods of annotation is considered to judge the fidelity of the annotation.

Companies names are anonymized to reduce bias. Industry/domain information about the content is marked to provide context to the annotator. For example, Merrill Lynch is marked as a banking company. We randomly select 600 articles and use Amazon Mechanical Turk for crowdsourcing the text annotation. After filtering using consistency check mentioned above, we obtain articles that have scores from at least two annotators. We further shortlist the articles only if at least two annotators have agreed whereby we obtain 500 annotated articles. Two annotators are said to have agreed on a given article for a particular dimension if the absolute difference between their scores is less than or equal to 1. For example, resolving mismatch such as high sincerity may mean a score of 4 to one annotator and 5 to another annotator. This normalizes the biases between annotators. We provide the inter-annotator agreement for each dimension in Table 3, which averages to 67.25%. We take the average of the score given by the annotators at this stage and use a static threshold of 3.0 to convert the score to a binary label, indicating whether a particular brand persona is evoked from the text or not.

4.2 Linguistic Features

We extract the following set of linguistic features from each article. We use the concepts proposed by Delin [6] to formulate a number of linguistic features, which aim to capture the trait of the underlying article in a more compact form.

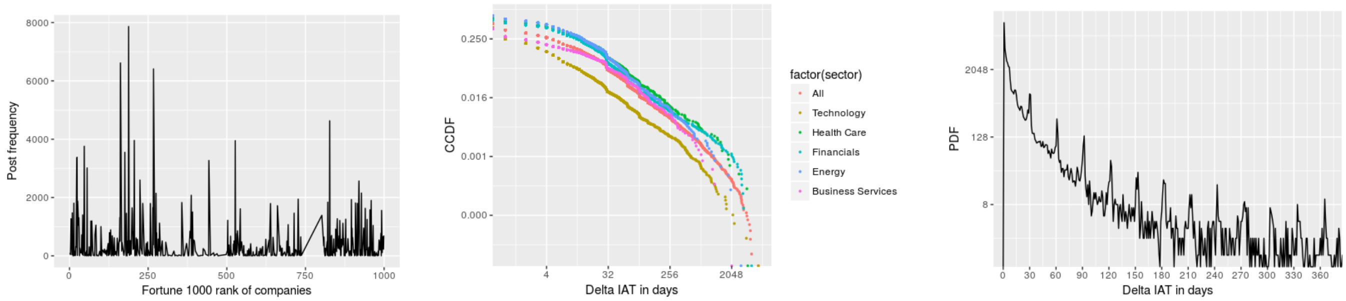

Figure 1: (left) Line plot of total posting volume per company. (middle) Heavy-tailed distribution of postings IAT over all companies and also a sector-wise trend. (right) Periodic peak pattern: the log-scale histogram of posting's IAT has peaks at 30-33 days interval

Trait	Positive	Negative	Agreement(in %)
sinc	433	67	71.41
exc	339	161	65.02
com	470	30	75.62
rug	190	310	63.50
soph	276	224	60.70

Table 3: Class distribution and inter–annotator agreement per dimension for HT

LIWC: Linguistic Inquiry and Word Count [23] is a dictionary of psycho-linguistics traits. This has been frequently used in social data analysis [26] and psychological trait extraction. We use the values returned by the commercial API version [3] of LIWC for a particular piece of text.
Term Frequency-Inverse Document Frequency (TF-IDF): A TF-IDF vector per document consisting of unigrams, bigrams, and trigrams after removing stop words is considered.
Contractions: These are shortened versions of phrases. Some examples are `we are` being replaced by `we're`, `is it not` being substituted by `isn't`. Contractions add a degree of informality and a conversational tone to the text.
Collocations: Collocations are word combinations that are known to occur frequently together. Examples include heavy rain, high temperature, etc. We use the Pearson Academic Collocations `cite` list which consists of 2,469 most frequent lexical collocations in written academic English, to form our dictionary of collocations.
Chains of reference: Chains of reference denote the brand's use of references to itself and closely associated elements. The key elements in this process are the noun phrases in the text, whose content and form can serve either to strongly evoke a brand, reinforce it, or not evoke it at all. Different kinds of relations that can hold between noun phrases and brand concepts as categorized by [6] have been summarized in Table 4. For our model, we use repetition, partial repetition, co-reference, and possessive inferrable as four different features.
Readability: This feature intuitively captures the ease of reading a given piece of text. The feature is based on the Flesch–Kincaid Readability Score [8]. The score considers the word length, sentence length, and the number of syllables per word. The higher the value of this score, the easier the content is for reading. It is calculated as follows:

$$Score = 206.835 - 1.015\frac{TotalWords}{TotalSentences} - 84.6\frac{TotalSyllables}{TotalWords}$$

[3]www.receptiviti.ai

Link	Definition	Example
Repetition	Repeating the full reference to the brand	Orange...Orange , Target...Target
Partial Repetition	A phrase contains a reference to the brand, but refers to something other than the brand concept	Orange...the Orange Service Promise , Target...Target Stores
Co-reference	Where a concept is reinforced by referring to it again, but not using a full descriptive noun phrase	Orange...We, Target... With us
Possessive inferrables	Where a link is created by referring to something that the brand has, does, or has given to the customer, using a possessive noun phrase	Orange...our network your phone, Target... our stores your cart

Table 4: Categorization of Chains of reference

4.3 Classification Model

We train separate classifiers for the five traits independently. The classification models, identical for each dimension. The classifier uses only human annotated data. The annotated data act as ground truth while the features discussed in the previous section are fed into a classifier. We consider several classification models which include Naive Bayes, Logistic Regression and Decision Tree, linear support vector machines classification models as well as ensemble algorithms like Random Forest and AdaBoost. During classification, we observe a large class imbalance in the human–annotated data, as summarized in Table 3. We use a data-level approach called SMOTE [5] to address the class-imbalance problem, which works by modifying the training set [24].

Trait	sincerity			excitement			competence			ruggedness			sophistication		
	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1
Naive Bayes	0.238	0.851	0.371	0.162	0.791	0.268	0.319	0.941	0.721	0.268	0.622	0.319	0.141	0.854	0.239
Logistic Regression	0.538	0.856	0.659	0.796	0.801	0.798	0.774	0.953	0.853	0.789	0.561	0.654	0.733	0.727	0.725
DecisionTree	0.774	0.871	0.819	0.661	0.742	0.698	0.921	0.954	0.937	0.568	0.536	0.549	0.652	0.671	0.66
RandomForest	0.838	0.865	0.85	0.72	0.795	0.754	0.953	0.939	**0.946**	0.532	0.641	0.575	0.623	0.737	0.673
AdaBoost	0.85	0.868	0.859	0.746	0.761	0.753	0.923	0.95	0.936	0.606	0.585	0.589	0.66	0.69	0.672
SVM (Linear)	0.912	0.861	**0.885**	0.832	0.801	**0.815**	0.919	0.943	0.931	0.773	0.57	**0.655**	0.751	0.707	**0.725**

Table 5: Performance comparison of the different binary classification algorithms for optimal classifer selection

Trait	sincerity			excitement			competence			ruggedness			sophistication			PMax	SpMax
	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1		
liwc (baseline)	0.912	0.861	0.885	0.832	0.801	0.815	0.919	0.943	0.931	0.773	0.57	**0.655**	0.751	0.707	**0.725**	0.409	0.372
liwc+ tfidf	0.988	0.867	0.923	0.9	0.787	**0.839**	0.998	0.94	**0.968**	0.527	0.573	0.545	0.726	0.704	0.707	0.407	0.429
tfidf*+ contractions	0.989	0.867	0.923	0.894	0.784	0.834	0.998	0.94	0.968	0.527	0.578	0.548	0.725	0.7	0.708	0.406	0.424
cont*+ collocations	0.989	0.867	0.923	0.897	0.787	0.837	0.998	0.94	0.968	0.522	0.579	0.545	0.722	0.708	0.709	0.405	0.428
coll*+ chainref	0.991	0.867	**0.925**	0.894	0.787	0.836	0.998	0.94	0.968	0.559	0.587	0.569	0.729	0.693	0.706	0.404	0.407
chainref*+ readability	0.989	0.868	0.924	0.862	0.805	0.837	0.998	0.94	0.968	0.59	0.6	0.592	0.722	0.718	0.72	0.406	0.419
Best features (FLCS)	0.991	0.867	0.925	0.9	0.787	0.839	0.998	0.94	0.968	0.773	0.57	0.655	0.751	0.707	0.725	**0.426**	**0.431**

Table 6: Comparing performance across different feature sets and forming the Final Linear Classifier Set (FLCS)

4.4 Metrics to test the classifier

We observe an uneven ratio of the number of positive and negative class data points across all the traits as evident in Table 3 and thus solely maximizing accuracy will tend to favor the class with more number of data points. We compare the performance of the different classification models by reporting the precision, recall and F1 score. As previously observed, there is a significant class imbalance. Therefore we use F1 score as a single metric for both selecting the optimal classification model and also for choosing the optimal feature set. We thus perform a 7-fold cross-validation strategy and compute our final score as the average of the obtained scores.

4.5 Results

We first perform experiments to select the best classification model. We then identify the set of feature sets which produce the best result, We further validate our methodology by correlating the trait-wise scores, and the ranks provided the annotators (explained in Section 4.1) over the metrics of Pearson coefficient and Spearman's rank correlation coefficient.

4.6 Best classification model

To determine which classifier among the alternatives performs best, we first need to choose a feature set that is already established in terms of the brand personality detection task. Xu et al. [27] use LIWC to characterize the factors contributing to shaping brand personality on social media. We, therefore, decide to use LIWC comprising of 64 categories as our feature set for determining FLC. We, therefore, train all the candidate supervised classification models. We observe that linear SVM reports the highest F1-score for all the brand personality traits except competence as shown in Table 5. We, therefore, choose linear SVM as our final classifier since it has the **highest average F1-score of 0.802**, combining all traits.

4.7 Feature addition

After the selection of linear SVM, we add different linguistic features as described in Section 4.2, on top of the established feature set of LIWC. We incrementally expand the feature sets and choose the optimal set of features which give the highest accuracy. We are able to achieve a very high F1-score for our brand personality consistency task for sincerity, excitement, and competence. For the remaining two traits, ruggedness and sophistication, we observe a good precision score of 0.773 and 0.751 respectively. If we consider the best set of features for each class, we are able to achieve an **a F1-score of 0.822**, which is an improvement over the (limited-feature-set) linear SVM by 2.49%.

We further validate our model by utilizing the ground-truth in which the annotators have provide the ranks of the five brand personality traits in order of their presence in the provided text as well as individual trait-wise score. For each text, we obtain individual trait-wise confidence score from their respective classifiers and create a ranking among the five traits. We then compute the Pearson correlation coefficient between the individual trait-wise scores provided by the annotators with the confidence score provided by the classifiers. We also compute the Spearman rank correlation coefficient between the ranks provided by the annotators and the rank that we obtain based on the confidence scores. We observe from Table 6 that the improvement (deterioration) due to feature addition highly correlate with both the Pearson and Spearman score. More importantly, we observe that the classifier set formed using the best feature set (which is different for different traits), also performs the best in both these two metrics (Pearson - 0.426, Spearman - 0.431). This provides an extra validation to our proposed methodology.

We call the set of five classifiers with the best performing features (different for each trait) model as FLCS (Final Linear Classifier Set), and we use it for annotating MT_{large} required for the brand consistency study, covered in the next section.

4.8 High Fidelity Points

We now use FLCS to classify the MT_{large} data and only select those data points which are classified with high confidence (≥ 0.095), to carry out our brand consistency study. We manually cross-check whether the selected dataset indeed confirms to the class to which it is annotated. A random checking of 50 data points yields just two errors. We name this dataset as MT_{high} which comprises of 93321 data points covering 536 companies. Here, we determine the points have high confidence based on the distance value from the decision boundary, which we obtain for each of the five traits for each text article. We now empirically determine a threshold value for each trait, above which we will tag the text with that particular trait to be present, else tag it to be absent. This acts as a sanity check for our high fidelity dataset. We finally construct the MT_{high} dataset, as having points which have at least one trait present. We now describe the different data collection steps that we perform until now in Table 7.

Dataset name	Number of posts	Number of companies	Collection strategy
MT_{large}	298112	643	Web scraping from official websites based on accept and deny keywords
HT	500	-	Randomly selected 600 points from MT_{large}, which satisfy strict annotation criteria
MT_{high}	93321	536	Subset of MT_{large}, which is annotated with high confidence by FLCS
MT_{time}	49833	242	Subset of MT_{high} having timestamp data
MT_{noTime}	43488	512	Subset of MT_{high} without having timestamp data

Table 7: Summary of the different data collection steps

5 BRAND CONSISTENCY ANALYSIS

We now perform a characterization study where we investigate how well a company exhibits and maintains its brand personality across the web content both across time and also over different content categories. As per our knowledge, this is the first study which investigates the notion of brand consistency computationally. To conduct the study, we first define the consistency between the two texts.

5.1 Brand Consistency - Definition

Consumers tend to interact with companies in a manner similar to humans, where they closely try to follow their inter-personal and social relationships. Therefore, when a company's action violates the relationship norms, consumers develop a more negative perception of the brand as compared to when the brand actions are consistent with those relationship norms [2]. InterBrand [4] - a global brand agency, considered consistency to be among the ten strengths of companies, responsible for sustainable growth of brands among other properties like clarity, commitment, governance, responsiveness, authenticity, relevance, differentiation, presence and engagement.

Formulation: As discussed each post can be represented by two 5-dimensional vectors - *label vector* and *rank vector*. Label vector stores the binary label of whether a trait is present or absent in the text article. Rank vector stores an order of precedence of the brand personality traits from the textual content computed based on its confidence score (higher is the distance from the decision boundary, higher is its confidence score). We first scale this distance value by subtracting the trait-specific threshold value (as mentioned in Section 4.8, before computing the order. We calculate the similarity between a post and the representative vectors (static post) of the respective company using two measures - *binLabelSim* and *rankLabelSim*. Based on their values, we categorize the consistency level into four categories as shown in Table 8. We only choose the most frequently occurring label and rank vector among all the static posts as the respective representative vectors (stands for the company's brand personality). Here, we observe that the static posts are highly consistent among themselves with the average pair–wise binLabelSim being 0.935 (std. deviation of 0.129) and rankVectorSim being 0.861 (std. deviation of 0.285). On the other hand, dynamic posts show an overall average value of binLabelSim as 0.65 and rankVectorSim as -0.03.

Similarity Measure: We now explain how we compute the similarity distance value between two texts, each of which is represented by the above-mentioned label vector and a rank vector.

binLabelSim: To compute the distance between two posts (static and dynamic), we consider their label vectors and use standard distance measures of hamming and Levenshtein distance and compute a composite score termed as *binLabelDist*, which is the simple average of their hamming and Levenshtein distance measures. *binLabelSim* is 1 - *binLabelDist*.

rankVectorSim: We use Pearson, Kendall tau and Spearman's rank correlation coefficient to compute the similarity between two such rank vectors. We calculate a composite score (*rankVectorSim*) which is simply the average of Pearson, Spearman and Kendall

[4] https://www.interbrand.com/best-brands/

tauś rank correlation scores. An important point to mention here is that Spearman and Kendall's tau compute correlation using the respective rank vectors, whereas, for computing Pearson correlation, we directly use the scaled confidence score from each of the trait-specific classifiers.

Consistency Levels: We empirically determine the conditions for the different brand consistency levels (see Table 8). The conditions are somewhat arbitrarily determined. From manual inspection, we have noticed that binLabelSim has a higher importance in reflecting the level of brand consistency. Accordingly, a strict ordering of binLabelSim is maintained while rankVectorSim is used as a secondary measure to ensure consistency.

Consistency Score: Given a set of posts whose *binLabelSim* and *rankVectorSim* is calculated, consistency score (ConsScr) is defined as the ratio of the number of consistent post(label 1-3) and the total number of posts within that time-frame. For example, say for Microsoft Corporation in temporal bin index 1, which is for first 3 months, we have 10 such dynamic web posts, with the following brand consistency level breakup - strongly consistent (0 out of 10), partially consistent (1), somewhat consistent (3) and not consistent (6 out of 10). We will then say that ConsScr is equal to 0.4 in this case. A ConsScr closer to 1.0 indicates a higher degree of consistency. We consider a given temporal bin for a company to be consistent if $ConsScr \geq 0.5$.

Brand consistency level	binLabelSim	rankVectorSim
Strongly consistent	≥ 0.8	≥ 0.6
Partially consistent	≥ 0.8	≥ 0.2
Somewhat consistent	≥ 0.5	≥ 0.6
Not consistent	Otherwise	Otherwise

Table 8: Conditions associated with different degrees of brand consistency

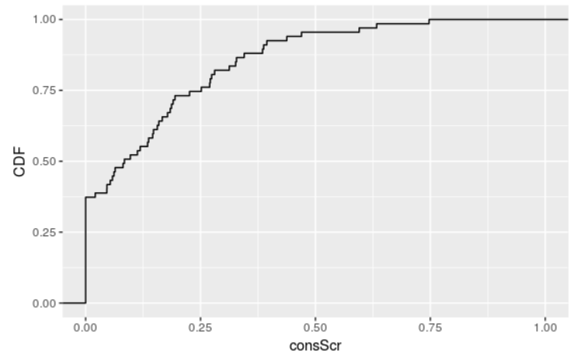

Figure 2: CDF showing consistency score across the companies.

Trait	Top 5 most prominent companies(number of posts)
sinc	Hospitality Properties Trust (159), Discover Financial Services (53), DaVita Inc. (41), Calpine Corporation (29), Darden Restaurants, Inc. (26)
exc	Microsoft Corporation (164), Tribune Media Company (42), Tutor Perini Corporation (29)
com	The Carlyle Group L.P. (67), CSX Corporation (51), Ally Financial Inc. (46), F5 Networks, Inc. (42), Vornado Realty Trust (37)
sop	Oceaneering International, Inc. (69), Tailored Brands, Inc. (62), Hawaiian Holdings, Inc. (45)

Table 9: Top 5 companies in terms of dominant trait in overall event promotions

5.2 Experimental Setup

Since brand consistency is an attribute of the company, we only consider the companies which satisfy a strict data requirement criteria. We only consider companies which have at least one static post with at least one trait present in it. This reduces the number of companies covered to 204. We also mention here that we have timestamp information for only 49833 points of MT_{high} (53.4%), which we call as MT_{time} and those not having time-stamp information as MT_{noTime}. We use both MT_{time} and MT_{noTime} for the studies. Figure. 2 shows consistency score vis-a-vis fraction of companies maintaining that consistency. As can be seen, only a very few companies can maintain high consistency highlighting the extent of the problem which we have already mentioned. Table 9 shows the top five companies displaying the maximum number of posts in each of the five categories.

5.3 Product promotion posts

The posts related to event or product promotions formed a significant portion of all posts. This corresponds to the description of *products & services* category, which is one of the seven reputation dimensions of the RepTrak Framework as described in the related work (Section 2). This category primarily contains information about the company's products and services, as well as about consumer satisfaction. We first construct a data subset by performing a lexicon based search to automate the identification of such posts. We check whether the following keywords - *event, promotions, promot, products, product-launch, announce, launch*, are present in web page URL. We thus obtain 3255 such data points which satisfy all the above criteria. We observe competence is the primary trait with product promotion followed by sincerity. Individual count where a trait major : sincerity - 839 (out of 3255), excitement - 462, competence - 1334, ruggedness - 0, sophistication - 620.

5.4 Top companies maintaining brand consistency

Here, we identify the Fortune 1000 companies that are best able to maintain brand consistency across their company postings. We measure this property in terms of the percentage of its company postings being strongly consistent and is outlined in Table 10. We also observe that all of these companies have a very high mean and very low standard deviation values of binLabelSim and rankVectorSim. We only consider companies that have at least 20 strongly consistent posts, since we are ranking with the percentage of such posts and not the absolute count of such posts.

spec domain	consistent posts (in %)	binLabel Sim mean	binLabel Sim sd	rankVector Sim mean	rankVector Sim sd
FTI Consulting, Inc.	100.0	0.9	0.1	0.68	0.08
Regis Corporation	54.0	0.88	0.1	0.67	0.06
Engility Holdings, Inc.	84.0	0.08	0.42	0.66	0.05
Caesars Entertainment Corporation	41.0	0.8	0.0	0.81	0.1
Prudential Financial, Inc.	23.0	0.91	0.1	0.68	0.08

Table 10: Top 5 companies with highest percentage of consistent dynamic posts

5.5 Top Companies maintaining temporal brand consistency

We now study the notion of brand consistency as a company attribute, rather than as a post attribute. Here, we quantify brand consistency score for a company in terms of the different brand consistency categories as mentioned in Table 8. Here we follow a temporal binning strategy where posts of 12 weeks are binned together. We mention the top 5 companies ranked in terms of the total number of such consistent temporal bins in Table 11. We see that only two of the top five companies overlap in the two sets shown by Tables 10 and 11.

spec domain	ConsScr mean	ConsScr sd	Total number of bins
Engility Holdings, Inc.	0.747	0.154	11
Regis Corporation	0.633	0.174	4
Principal Financial Group, Inc.	0.595	0.152	4
Westlake Chemical Corporation	0.47	0.316	11
Capital One Financial Corporation	0.438	0.241	45

Table 11: Top 5 companies with with temporal consistency score across bins

5.6 Top-ranked company vis-a-vis brand consistency

We consider the companies ranked within 150 among the Fortune 1000 companies as the top-ranked companies and those companies between the rank of 850 and 1000 as our lower ranked companies. We only consider companies that have at least 25 dynamic web pages, thus having a strict minimum data requirement. We are thus left with 18 top-ranked companies and 20 bottom-ranked companies for studying this research question. Here, we consider temporal bins of a duration of 6 months instead of the previous 12 weeks, since the number of data points per bin was observed to be very sparse. The first 5 top companies in terms of dynamic posts count - Microsoft Corporation (5365), Bank of America Corporation (467), Intel Corporation (294), Capital One Financial Corporation (282), Starbucks Corporation (143); similarly the first 5 bottom-ranked companies are - Red Hat, Inc. (732), Autodesk, Inc. (270), Engility Holdings, Inc. (225), Akamai Technologies, Inc. (180), Overstock.com, Inc. (128).

We observe that the top-ranked Fortune 1000 companies can maintain a higher average consistency score as compared to the bottom-ranked companies for the first 12 months, after which its consistency score drops to a score which is equal to the bottom-ranked companies (see Figure 3). On the other hand, bottom-ranked companies maintain a low, consistent score throughout the observed period of 2 years.

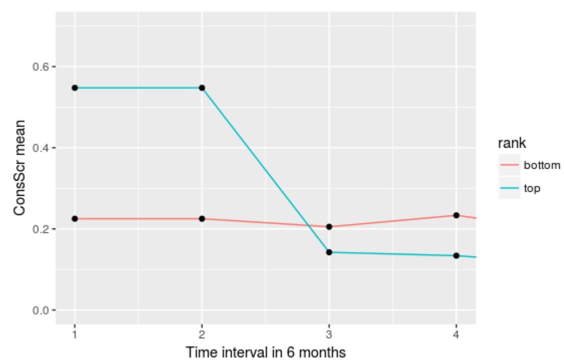

Figure 3: Consistency mean over temporal bins of duration 6 months comparing top and bottom ranked companies

6 CONCLUSION

To the best of our knowledge, this is one of the first attempts towards quantifying brand personality from the text content of an organization's official website. We launch a major crawling activity and are able to collect 298112 web page content covering 643 Fortune 1000 companies. We classify the dataset of each company into static content and dynamic posts and undertake a rigorous approach to find the timestamp of the dynamic posts. Further to it, we annotate a random sample of the data (around 600 data points); we undertake double checking measure whereby the same data points are annotated through labeling as well as ranking. We build five independent classifiers (linear SVM) one for each trait and finally optimize the feature set for each of them; the final classifier set is called FLCS. FLCS provides branding information to each of the text - the classification of the high confidence points is almost 100% correct. With the dataset thus annotated, we study the brand characteristics of companies. For that, we define several metrics and determine four levels of consistency. We discover companies which post consistently and find that more wealthy companies are better at maintaining consistency.

Limitations: We only consider textual content of web articles posted by the companies themselves and do not cover any form of user-generated content regarding the companies. Another important limitation is that we do not cover the visual or content-independent aspects of a brand style guide like color, typography and positioning of different sections and headers of a brand website.

Future Work: In the current work, we develop independent classifiers for each trait. However, it might be possible that one trait (weakly) implies one or more of the others; thus jointly learning all the traits together would be an important future work. We will use the insights derived from this work dealing with document-level text classification and move onto finer granularity like sentence level and identify the most contributing sentences towards the expression of a brand. We will further extend our work to develop a helper tool for the content writers and brand managers, which sentences should be modified for making the text articles more consistent with the targeted brand personality. These would be our next future endeavors.

ACKNOWLEDGMENTS

The first author is currently supported by the Junior Research fellowship from the project titled "IoTDiff: Service Differentiation for Personal Hubs", which is sponsored by Science & Engineering Research Board, New Delhi. A part of the work was carried out during the summer internship at Big Data Experience Lab, Adobe Research, Bangalore, India.

REFERENCES

[1] Jennifer L Aaker. 1997. Dimensions of brand personality. *Journal of marketing research* (1997), 347–356.

[2] Pankaj Aggarwal. 2004. The effects of brand relationship norms on consumer attitudes and behavior. *Journal of consumer research* 31, 1 (2004), 87–101.

[3] Enrique Amigó, Jorge Carrillo-de Albornoz, Irina Chugur, Adolfo Corujo, Julio Gonzalo, Edgar Meij, Maarten de Rijke, and Damiano Spina. 2014. Overview of RepLab 2014: Author Profiling and Reputation Dimensions for Online Reputation Management. In *Information Access Evaluation. Multilinguality, Multimodality, and Interaction*, Evangelos Kanoulas, Mihai Lupu, Paul Clough, Mark Sanderson, Mark Hall, Allan Hanbury, and Elaine Toms (Eds.). Springer International Publishing, Cham, 307–322.

[4] Enrique Amigó, Jorge Carrillo-de Albornoz, Irina Chugur, Adolfo Corujo, Julio Gonzalo, Edgar Meij, Maarten de Rijke, and Damiano Spina. 2014. Overview of RepLab 2014: Author Profiling and Reputation Dimensions for Online Reputation Management. In *Information Access Evaluation. Multilinguality, Multimodality, and Interaction*, Evangelos Kanoulas, Mihai Lupu, Paul Clough, Mark Sanderson, Mark Hall, Allan Hanbury, and Elaine Toms (Eds.). Springer International Publishing, Cham, 307–322.

[5] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. 2002. SMOTE: Synthetic Minority Over-sampling Technique. *J. Artif. Int. Res.* 16, 1 (June 2002), 321–357. http://dl.acm.org/citation.cfm?id=1622407.1622416

[6] Judy Delin. 2007. Brand Tone of Voice. *Journal of Applied Linguistics* 2, 1 (2007). https://journals.equinoxpub.com/index.php/JAL/article/view/1471

[7] Alecia Douglas, Juline Mills, and Raphael Kavanaugh. 2007. Exploring the use of emotional features at romantic destination websites. *Information and Communication Technologies in Tourism 2007* (2007), 331–340.

[8] J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. 1975. *Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel.* Technical Report. DTIC Document.

[9] Shibamouli Lahiri. 2015. SQUINKY! A Corpus of Sentence-level Formality, Informativeness, and Implicature. *arXiv preprint arXiv:1506.02306* (2015).

[10] Wang Ling, Tiago Luís, Luís Marujo, Ramón Fernandez Astudillo, Silvio Amir, Chris Dyer, Alan W Black, and Isabel Trancoso. 2015. Finding function in form: Compositional character models for open vocabulary word representation. *arXiv preprint arXiv:1508.02096* (2015).

[11] Fei Liu, Julien Perez, and Scott Nowson. 2017. A Language-independent and Compositional Model for Personality Trait Recognition from Short Texts. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, Vol. 1. 754–764.

[12] Liu Liu, Daria Dzyabura, and Natalie Mizik. 2018. Visual listening in: Extracting brand image portrayed on social media. In *Workshops at the Thirty-Second AAAI Conference on Artificial Intelligence*.

[13] X. Liu, A. Xu, L. Gou, H. Liu, R. Akkiraju, and H. Shen. 2016. SocialBrands: Visual analysis of public perceptions of brands on social media. In *2016 IEEE Conference on Visual Analytics Science and Technology (VAST)*. 71–80. https://doi.org/10.1109/VAST.2016.7883513

[14] Zhe Liu, Anbang Xu, Yi Wang, Jerald Schoudt, Jalal Mahmud, and Rama Akkiraju. 2017. Does Personality Matter?: A Study of Personality and Situational Effects on Consumer Behavior. In *Proceedings of the 28th ACM Conference on Hypertext and Social Media (HT '17)*. ACM, New York, NY, USA, 185–193. https://doi.org/10.1145/3078714.3078733

[15] Natalia Maehle, Cele Otnes, and Magne Supphellen. 2011. Consumers' perceptions of the dimensions of brand personality. *Journal of Consumer Behaviour* 10, 5 (2011), 290–303.

[16] François Mairesse, Marilyn A Walker, Matthias R Mehl, and Roger K Moore. 2007. Using linguistic cues for the automatic recognition of personality in conversation and text. *Journal of artificial intelligence research* 30 (2007), 457–500.

[17] Brigitte Müller and Jean-Louis Chandon. 2003. The impact of visiting a brand website on brand personality. *Electronic Markets* 13, 3 (2003), 210–221.

[18] Ellie Pavlick and Joel Tetreault. 2016. An empirical analysis of formality in online communication. *Transactions of the Association of Computational Linguistics* 4, 1 (2016), 61–74.

[19] Joseph T. Plummer. 2000. How Personality Makes a Difference. *Journal of Advertising Research* 40, 6 (2000), 79–83. https://doi.org/10.2501/JAR-40-6-79-83 arXiv:http://www.journalofadvertisingresearch.com/content/40/6/79.full.pdf

[20] Damiano Spina, Julio Gonzalo, and Enrique Amigó. 2014. Learning Similarity Functions for Topic Detection in Online Reputation Monitoring. In *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval (SIGIR '14)*. ACM, New York, NY, USA, 527–536. https://doi.org/10.1145/2600428.2609621

[21] Ming-Hsiang Su, Chung-Hsien Wu, and Yu-Ting Zheng. 2016. Exploiting turn-taking temporal evolution for personality trait perception in dyadic conversations. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 24, 4 (2016), 733–744.

[22] X. Sun, B. Liu, J. Cao, J. Luo, and X. Shen. 2018. Who Am I? Personality Detection Based on Deep Learning for Texts. In *2018 IEEE International Conference on Communications (ICC)*. 1–6. https://doi.org/10.1109/ICC.2018.8422105

[23] Yla R. Tausczik and James W. Pennebaker. 2010. The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. (2010). http://homepage.psy.utexas.edu/homepage/students/Tausczik/Yla/index.html

[24] Jing Wang and Min-Ling Zhang. 2018. Towards Mitigating the Class-Imbalance Problem for Partial Label Learning. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2018, London, UK, August 19-23, 2018*. 2427–2436. https://doi.org/10.1145/3219819.3220008

[25] Honghao Wei, Fuzheng Zhang, Nicholas Jing Yuan, Chuan Cao, Hao Fu, Xing Xie, Yong Rui, and Wei-Ying Ma. 2017. Beyond the Words: Predicting User Personality from Heterogeneous Information. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining (WSDM '17)*. ACM, New York, NY, USA, 305–314. https://doi.org/10.1145/3018661.3018717

[26] Anbang Xu and Brian Bailey. 2012. What do you think?: a case study of benefit, expectation, and interaction in a large online critique community. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*. ACM, 295–304.

[27] Anbang Xu, Haibin Liu, Liang Gou, Rama Akkiraju, Jalal Mahmud, Vibha Sinha, Yuheng Hu, and Mu Qiao. 2016. Predicting Perceived Brand Personality with Social Media. In *Proceedings of the Tenth International Conference on Web and Social Media, Cologne, Germany, May 17-20, 2016*. 436–445. http://www.aaai.org/ocs/index.php/ICWSM/ICWSM16/paper/view/13078

[28] Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 1480–1489.