WebSci 2019

Published in 11th ACM Conference on Web Science,
Northeastern University, Boston, USA

# Understanding Brand consistency from Web content

**Soumyadeep Roy**, Niloy Ganguly, Shamik Sural (IIT Kharagpur, India)
Niyati Chhaya, **Anandhavelu Natarajan** (Adobe Big Data Experience Lab, India)

INDIA HCI 2019

"Out-of-India" Exhibition track at India HCI 2019

# Brand Personality: Consumer's perception of a company

- People talk (communicate) with other people. Get to know the personality of the person
  - outgoing, talkative, and energetic versus reserved and solitary?
  - sensitive and nervous versus secure and confident?

# Brand Personality: Consumer's perception of a company

- People talk (communicate) with other people. Get to know the personality of the person
  - outgoing, talkative, and energetic versus reserved and solitary?
  - sensitive and nervous versus secure and confident?

- Similarly, companies communicate with their consumers (B2C setting). Consumers form a perception of the company
  - advertisements, tweets, blog posts, media releases, product launch events

# What is Brand Personality ?

- Component of brand image, used in marketing campaigns
  - Understanding of consumer choice

- Aaker(1997) formalizes the concept of brand personality into five dimensions

| sincerity | excitement | competence | ruggedness | sophistication |
|---|---|---|---|---|
| Down-to-earth<br>Honest<br>Wholesome<br>Cheerful | Daring<br>Spirited<br>Imaginative<br>Up-to-date | Reliable<br>Intelligent<br>Successful | Outdoorsy<br>Tough | Upper class<br>Charming |

Aaker, J. L. 1997. Dimensions of brand personality. Journal of marketing research 347–356.

# Brand Consistency - Maintaining Brand personality over time and content

Rise and adoption of digital marketing to keep up the consumer engagement

Companies have to produce online content more frequently


Image source: The Balance

# Problem Statement

- Develop independent binary classification models for each dimension to predict whether a brand personality is present or absent from the text
  - Previous studies focus on user-generated content like tweets. Focus at content-creation time
- Formulate and analyze the notion of brand consistency on a large scale
- Develop a content-monitoring system to identify inconsistent posts and explain the contributory factors
- Develop recommendation systems for improving brand consistency

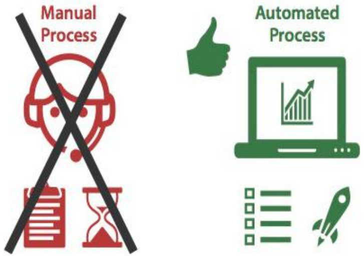
# Problem Statement: current task details

- Develop independent binary classification models for each dimension to predict whether a brand personality is present or absent from the text
  - Previous studies focus on user-generated content like tweets. Focus at content-creation time
- Formulate and analyze the notion of brand consistency on a large scale
- Develop a content-monitoring system to identify inconsistent posts and explain the contributory factors
- Develop recommendation systems for improving brand consistency

# Research challenges of Brand consistency

- Monitoring and maintaining brand consistency on a large scale is difficult and require costly human experts
    - Brand manager, content writers use **brand style guide**
    - No established metric


- Well tagged datasets for brand personality is not available



# Key takeaways and contributions

- Develop a supervised classification model with a high F1-score of 0.822
- Collect ~**300K** web pages covering ~**650 Fortune 1000** companies, posted between **January 2000** and **September 2017**
- Investigate how well a company maintains its brand personality across time and over different content categories.
  - Companies that post consistently and are higher ranked are better at maintaining brand consistency

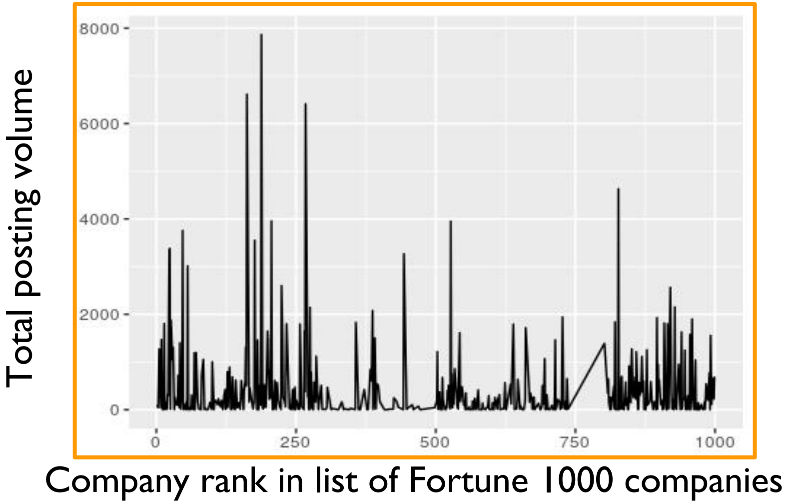
# Dataset

- Consider 2017 Fortune 1000 companies from their official websites through data crawling
  - Only consider company web pages that are directed towards the customers - about the company, media releases, blogs and communications

- Consider postings between January, 2000 and September, 2017
- Extract timestamp data from page url and also from text content(around 50%)
  - 75% contain day level information while the remaining contain only year level information

# Basic observations - 1

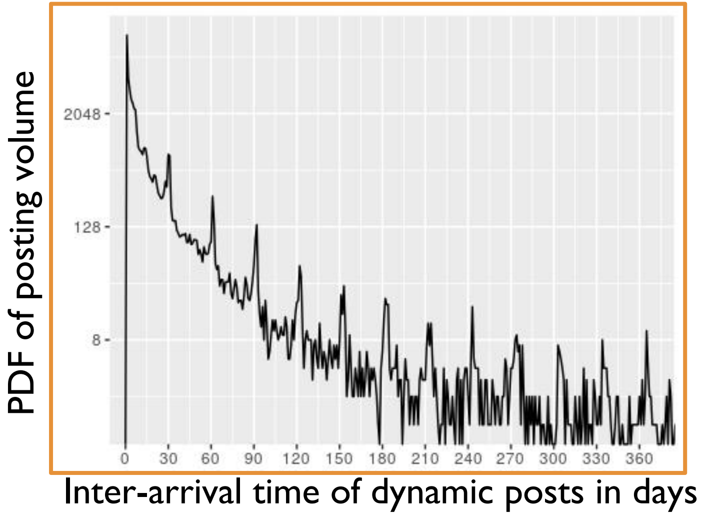
- Volume of posts is roughly similar in both top as well as bottom ranked companies




More number of spikes(indicates companies who post way more than average) occur among the top-ranked companies

# Basic observations - 2

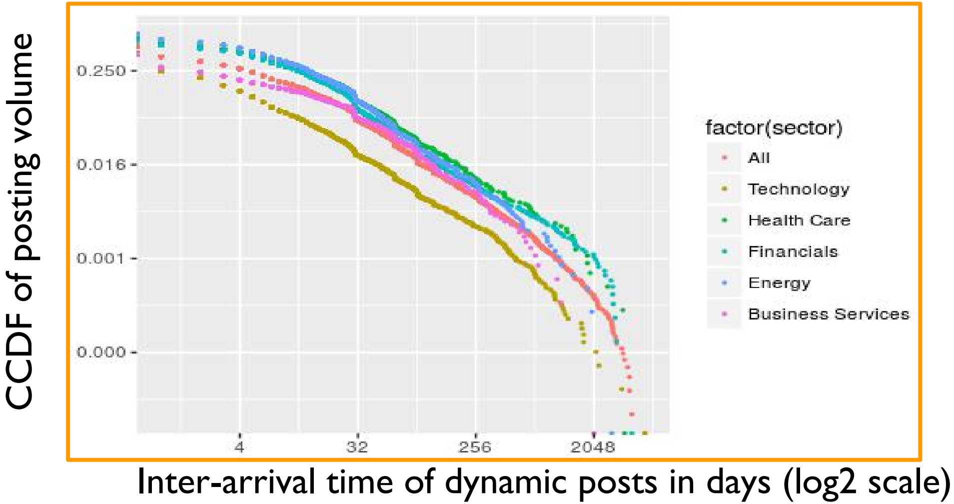
- We observe that these peaks are composed of around 78% web pages that are posted at the month end. The highest number belonged to post type ‘news’.
  - Dynamic posts - Maintaining continuous engagement with c like blogs, news, media or press releases




Posting volume show peaks at an interval of 30-33 days in terms of inter-arrival time

# Basic observations - 3

- Top 5 sectors in terms of number of posts display heavy tail behavior
    - technology(48219), financials(11739), energy(4915), healthcare(4685), business services(3747)




Inter-arrival time postings follow a heavy-tail pattern. Similar across different sectors

# Methodology workflow

Select best performing classifier using HT

Brand personality Linguistic features

$MT_{high}$
High Fidelity Data
(~93K articles from 536 companies)

Consistency formulation & characterization study

HT
Randomly selected 600 articles
Crowdsourced annotation
500 data points

$MT_{large}$
Data creation and analysis
(300K webpages from 643 companies )

# Brand SVM Classifier (**BrSVM**; Feature : **LIWC**)

- Traditional classifiers : *Naive Bayes, Logistic Regression, Decision Tree, Random Forest, AdaBoost, Linear SVM*
- Feature set used: *LIWC*

- **Feature Set selection :** We incrementally expand the feature set and add linguistic features on top of LIWC
  - LIWC > tfidf > contractions > collocations > chains of reference > flesch readability ease
  - Different feature sets are optimal for the different brand personality traits

# Brand SVM Classifier (**BrSVM**; Feature : **LIWC++**)

**Exisiting**
LIWC(Xu et. al, 2016)

**Proposed**
- TF-IDF
- Contractions
- Collocations
- Chains of Reference
- Readability

- **Contractions -** Adds informality and conversational tone; Ex : isn't, we're
- **Collocations -** Frequently occurring word combinations from '*Pearson Academic Collocation List*'; Ex : very good, extremely good, big house
- **Readability -** Computed as Flesch's Readability Ease. Depends on word length, sentence length, # syllables per word
- **Chains of reference -** Use of reference to oneself and alike entities(noun phrases). Repetition, partial repetition, coreference, possessive inferrables

A. Xu, H. Liu, L. Gou, R. Akkiraju, J. Mahmud, V. Sinha, Y. Hu, and M. Qiao. 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., pages 436–445, 2016.

# Experimental Setup - First Level Classifier

- Annotation : HT data, 500 web articles manually annotated and validated
  - Trait ranks also provided by annotators in terms of degree of presence
  - Company names blinded with the ‘Sector’ the company it belongs to.
  - Inter-annotator agreement : 67.25%
- Traditional classifiers : Naive Bayes, Logistic Regression, Decision Tree, Random Forest, AdaBoost, Linear SVM
- Feature set used : LIWC (Xu et al. 2016)
- 7-fold validation results ( 6/7th for training and 1/7th for testing on HT data )

# BrSVM - Best classification model

| Trait | sincerity | excitement | competence | ruggedness | sophistication |
|---|---|---|---|---|---|
| Naive bayes | 0.371 | 0.268 | 0.721 | 0.319 | 0.239 |
| Logistic Regression | 0.659 | 0.798 | 0.853 | 0.654 | 0.725 |
| Decision Tree | 0.819 | 0.698 | 0.937 | 0.549 | 0.66 |
| Random Forest | 0.85 | 0.754 | **0.946** | 0.575 | 0.673 |
| AdaBoost | 0.859 | 0.753 | 0.936 | 0.589 | 0.672 |
| SVM (Linear) | **0.885** | **0.815** | 0.931 | **0.655** | **0.725** |

Our Linear SVM model(BrSVM) is able to achieve a F1-score of 0.822

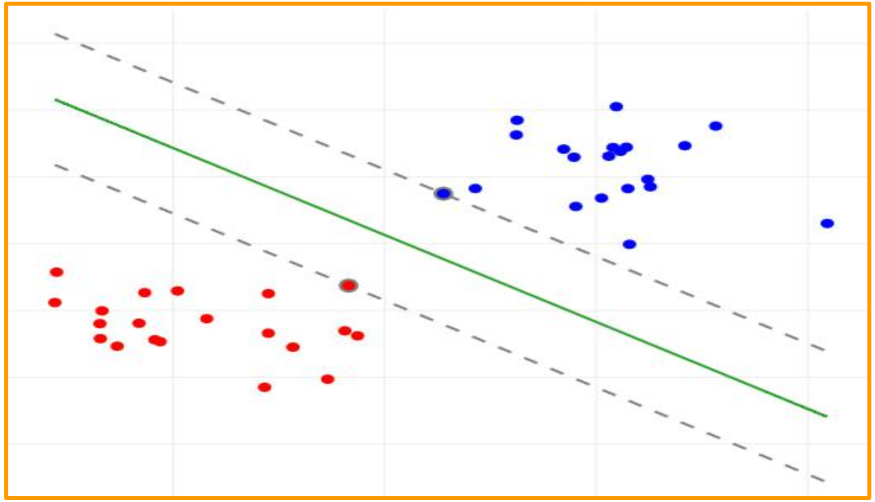
# BrSVM Feature Set Selection

| Feature sets | sincerity | excitement | competence | ruggedness | sophistication |
|---|---|---|---|---|---|
| LIWC (baseline) | 0.885 | 0.815 | 0.931 | **0.655** | **0.725** |
| tfidf* | 0.923 | **0.839** | **0.968** | 0.545 | 0.707 |
| contractions* | 0.923 | 0.834 | 0.968 | 0.548 | 0.708 |
| collocations* | 0.923 | 0.837 | 0.968 | 0.545 | 0.709 |
| chainref* | **0.925** | 0.836 | 0.968 | 0.569 | 0.706 |
| Best features | 0.925 | 0.839 | 0.968 | 0.655 | 0.725 |

We observe that the optimal feature set is different for each brand personality trait

# High fidelity data points

- FLCS used to classify $MT_{large}$ and only select that are classified with **high confidence(>= 0.095)**, which forms $MT_{high}$
- Threshold for each trait above for final output (present or absent)



# Dataset overview

| Dataset name | Total posts | Total 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 |

# Type of company posts - static and dynamic

- Static pages explicitly defines the brand which a company stands for
  - mission, vision and core values
  - **Static keywords:** introduction, about, commitment, people, vision, strength, history, approach, benefits
- Dynamic web pages comprises of content that is for continually engaging with the audience
  - blogs, news, media or press releases

# Brand consistency formulation

| sinc | exc | com | rug | sop |
|---|---|---|---|---|
| 1 | 1 | 0 | 0 | 0 |
| 2 | 1 | 3 | 4 | 5 |

- **Representation of a post :**
  - **Label vector -** Stores the binary label of whether a trait is present or absent in the text
  - **Rank vector -** Stores an order of precedence of traits based on confidence score

# Brand consistency formulation

| sinc | exc | com | rug | sop |
|---|---|---|---|---|
| 1 | 1 | 0 | 0 | 0 |
| 2 | 1 | 3 | 4 | 5 |

- Representation of a post :
  - **Label vector** - Stores the binary label of whether a trait is present or absent in the text
  - **Rank vector** - Stores an order of precedence of traits based on confidence score
- **Similarity measure :** We calculate the similarity between a dynamic post and the representative vectors(static post) of the company
  - **binLabelSim** - Levenshtein distance
  - **rankVectorSim** - Mean of Pearson, Kendall tau and Spearman rank correlation coefficient
- Formulate four brand consistency levels, out of which the lowest level is '**not consistent**'

# Top companies maintaining brand consistency

- Ranking based on the percentage of a company's posting being strongly consistent
- Ranking based on *ConScr mean value* for companies which have at least 20 strongly consistent temporal bins
- Observe that these companies have high mean and low standard deviation values of *binLabelSim* and *rankVectorSim*
- Uses **$MT_{notime}$ data**

# Top companies maintaining temporal brand consistency

- Notion at a company-level rather than as a post attribute
- Follow a temporal binning strategy where posts of 12 weeks are binned together
- Rank the companies in terms of highest temporal consistency score ( *Equal to average ConsScr value across all the temporal bins having at least 3 posts* )
- Use **$MT_{time}$ data**

Only few (5%) of companies are able to maintain high temporal consistency score

# Product promotion posts

- Corresponds to "products and services" category of the RepTrack framework
- Obtain **3255 articles** by performing a lexicon based search
- Competence is the primary trait with product promotion followed by sincerity

| Trait | Companies in descending order |
|---|---|
| sincerity | Hospitality Properties Trust (159), Discover Financial Services (53) |
| excitement | Microsoft Corporation (164), Tribune Media Company (42) |
| competence | The Carlyle Group L.P. (67), CSX Corporation (51) |
| sophistication | Oceaneering International, Inc. (69), Tailored Brands, Inc. (62) |

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

- Top-ranked companies : Within rank 1 - 150; Lead to 18 companies and Bottom-ranked companies : Within rank 850 - 1000; Lead to 20 companies

- Top ranked companies can maintain a higher company-level consistency score, on average, for the first 12 months than the bottom-ranked companies

# Limitations

- Only consider textual content of a web page and do not cover any user-generated content regarding the companies
- Do not consider other aspects of a brand style guide like color, typography, positioning of headers and website sections

# Limitations - Feedback required !!!!

- Only consider textual content of a web page and do not cover any user-generated content regarding the companies
- Do not consider other aspects of a brand style guide like color, typography, positioning of headers and website sections
- Features/meta-data that should be collected from a webpage to completely capture user perception
  - Scalable : data annotation is expensive
  - Evaluation metric for capturing perception, more experiments
  - Cross-checking, Error analysis

# Future Work

- Jointly learning the five brand personality traits, instead of independent classifiers
  - One brand personality trait may weakly imply another trait (Ex. competence with excitement)
  - Further improve the classifier performance using deep learning techniques
- Investigate at a sentence-level instead of document-level brand consistency score
- Given a not consistent web article, develop a helper tool targeting brand managers and content writers, to identify the sentences that needs to be modified

# Conclusion

- This is the first attempt to quantify brand personality from the text content of an organization's official website. Our proposed classification model, BrSVM achieves an **F1 score of 0.822**
- Collected around **300K** web page content covering around **650** Fortune 1000 companies and form an automatically annotated set, $MT_{high}$ containing very highly confident points
- We study the brand characteristics of a company and observe that companies that post consistently and are higher ranked are better at maintaining brand consistency

# Acknowledgements


- I am a part of the Complex Networks Research Group (CNeRG), IIT Kharagpur
  - Research group website: **http://www.cnergres.iitkgp.ac.in/**
- I am supported by the Senior Research Fellowship from Project IoTDiff, sponsored by Science and Engineering Research Board, New Delhi
  - Email id: **soumyadeep.roy9@iitkgp.ac.in**
  - Github username: **roysoumya**
- Conference travel and conference registration is provided by IIT Kharagpur Institute Funding and CNeRG Travel Grant

# References - Computing


- L. Liu, D. Dzyabura, and N. Mizik. Visual listening in: Extracting brand image portrayed on social media. In Workshops at the AAAI, 2018
- Z. Liu, A. Xu, Y. Wang, J. Schoudt, J. Mahmud, and R. Akkiraju. Does personality matter?: A study of personality and situational effects on consumer behavior. In Pr. ACM Conference on Hypertext and Social Media, HT '17, pages 185–193
- Anbang Xu, et al. 2016. Predicting Perceived Brand Personality with Social Media. ICWSM, 2016. 436–445
- D. Spina, J. Gonzalo, and E. Amigó. Learning similarity functions for topic detection in online reputation monitoring. In SIGIR 2014, pp. 527–536

# References - Consumer psychology


- Jennifer L Aaker. *Dimensions of brand personality.* Journal of marketing research, pp. 347-356, 1997
- 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
- J. Delin. *Brand tone of voice.* Journal of Applied Linguistics, 2(1), 2007

# Thanks for listening

Paper and slides: http://datanalytics101.com/

Please contact me to provide comment and feedback

soumyadeep.roy9@iitkgp.ac.in, @gmail.com

@roysoumya1

# Brand Consistency - Maintaining Brand personality over time and content

- In the era of digital marketing, organizations need to create a lot of online content to keep up the engagement with their audiences

- Organizations tends to maintain a *consistent perception among the customers* (brand personality) over time and across content categories
  - Generate trust and retain more customers
  - Current strategies for maintaining brand image over time is of qualitative nature, but no quantitative measure exists