Topic Modelling on Consumer-Generated Corpora

Automating Customer Relations Management with Text Analytics

By

Dayta, Dominic; Barrios, Erniel

School of Statistics, UP Diliman

Presented By:

Dayta, Dominic

MS (Statistics), UPSS

14th NCS National Convention on Statistics

1-3 October 2019 | Crowne Plaza Manila Galleria

Organized by the Philippine Statistical System

Spearheaded by the Philippine Statistics Authority

dbdayta@up.edu.ph

Agenda

1. Background
2. Topic Modelling, Latent Dirichlet Allocation
3. Initial Look Into The Data
4. Results
5. Current Challenges
6. Generalizations

dbdayta@up.edu.ph

Background | Big Data and Customer Feedback

- Service-oriented companies (e.g. restaurants, retailers) rely heavily on feedback provided by their customers.
- Most (if not all) of these feedback come in the form of text: emails, text messages, conversation transcripts (for called-in feedback).
- Two difficulties:
  - Classifying and prioritizing highly manual, prone to error
  - Statistical analysis limited to what can be transferred to structured form

dbdayta@up.edu.ph

Background | Big Data and Customer Feedback

Case Study [Company XYZ]*

- Feedback received is received by agents at call center partner
- Agents manually read through verbatim feedback received and classify them by: (1) business area concerned, (2) level of urgency
- Analysis team picks up structured data produced by agents for use in reports

*For legal concerns, the name of the company has been omitted. In subsequent slides, specific brand designations and product terminologies will be replaced with dummy tokens.

dbdayta@up.edu.ph

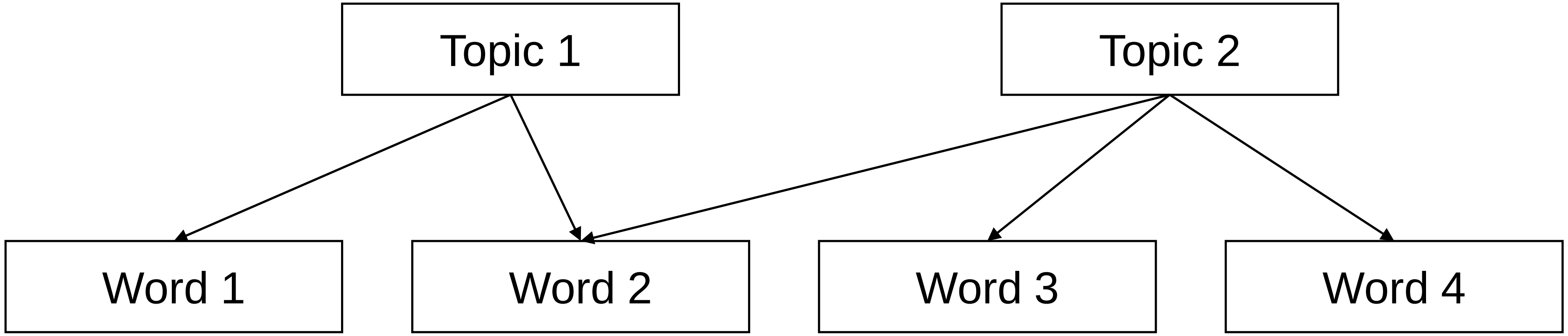
Topic Modelling | Making a Model that Reads

- An emerging field in analysis of unstructured big data, part of the bigger field of text analytics or natural language processing
- Concerns methods that attempt to summarize or classify documents based on word patterns and associations

dbdayta@up.edu.ph

Topic Modelling | Making a Model that Reads

Topic modelling treats document features (words) as observable variables that manifest under the influence of some higher, unobserved feature (topics) – as latent variables.



dbdayta@up.edu.ph

Topic Modelling | Making a Model that Reads

The test of a first-rate intelligence is the ability to hold two opposed ideas in the mind at the same time, and retain the ability to function. (The Crack-Up, F. Scott Fitzgerald)

Possible topic: ability, skill

dbdayta@up.edu.ph

Topic Modelling | Making a Model that Reads

The test of a first-rate intelligence is the ability to hold two opposed ideas in the mind at the same time, and retain the ability to function. (The Crack-Up, F. Scott Fitzgerald)

Possible topic: measure, ranking

dbdayta@up.edu.ph

Topic Modelling | Making a Model that Reads

- Topic modelling aims to construct "topic structures", i.e. distributions defining the probability or likelihood of certain words appearing in a document having a certain topic.

dbdayta@up.edu.ph

Topic Modelling | The Vector Space Model
(Xing, 2012)

Approach natural language as vectors:

- Documents are interpreted as vectors of word scores (counts, inv. freq, tf-idf)
- In this approach, order is ignored.
- Also called a “bag of words” approach

The test of a first-rate intelligence is the ability to hold two opposed ideas in the mind at the same time, and retain the ability to function.

Word	Count
The	5
Ability	2
To	2
A	1
And	1
At	1
First	1
…	…

dbdayta@up.edu.ph

Topic Modelling | Latent Dirichlet Allocation
(Blei et al, JMLR 2003)

- Developed out of two previous topic modelling procedures: Latent Semantic Indexing (LSI) and Probabilistic Latent Semantic Indexing (pLSI)
- Produces a *generative model* that describes how documents arise from multiple topics
- Bayesian in flavor

dbdayta@up.edu.ph

Topic Modelling | Latent Dirichlet Allocation

(Blei et al, JMLR 2003)

Some assumptions:

- Topics are independent of each other

  (i.e., the expression of one topic cannot magnify or preclude the expression of another topic within the same document)

- Documents exhibit multiple topics
- A topic is a distribution over a fixed set of vocabulary

  At the same time, this is assumed to come *before* the document. A document arises out of the admixing of topics.

- The number of possible topics is known in advance

dbdayta@up.edu.ph

Topic Modelling | Latent Dirichlet Allocation
(Blei et al, JMLR 2003)

Modelling topics using LDA:

- Consider a word as a basic discrete unit. In a corpus, there are W unique words, $w_i: i \in \{1,2,\ldots,W\}$
- A document is a corresponding sequence of words, denote as $\boldsymbol{w}_d = \{w_{1,d}, w_{2,d}, w_{3,d}, \ldots, w_{N,d}\}$ from the first word w_1 to the last w_N
- A corpus is a collection of documents $\boldsymbol{D} = \{\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_M\}$

dbdayta@up.edu.ph

Topic Modelling | Latent Dirichlet Allocation
(Blei et al, JMLR 2003)

Suppose the following: K = number of topics, α a positive K-dimensional vector, and η a scalar.

- For each topic, draw a distribution of words $\beta_K \sim Dirichlet(\eta)$
- For each document,
  - Draw a vector of topic proportions $\theta_d \sim Dirichlet(\alpha)$
  - For each word draw a topic assignment $Z_{d,n} \sim Multinomial(\theta_d)$
  - Draw a word $w_{d,n} \sim Multinomial\left(\beta_{z_d,n}\right)$

dbdayta@up.edu.ph

Topic Modelling | Latent Dirichlet Allocation
(Blei et al, JMLR 2003)

Two particularly useful vectors produced by LDA

- θ_d gives the proportion of topics inside document d
- β_k gives the probability of each word for topic k

- These values are estimated by integrating over the resulting posterior probability distribution.
- Complexity of the distribution leads to usage of numerical methods, popularly **Gibbs Sampling**.

dbdayta@up.edu.ph

Topic Modelling | Latent Dirichlet Allocation

(Blei et al, JMLR 2003)

How many topics?

- The number of topics affects the overall interpretability of the model
- Cross-validate!
- Due to the probabilistic nature of the method, a likelihood can be computed given a set number of topics. Choose the value that yields the highest likelihood.

dbdayta@up.edu.ph

Consumer Data | First Looks

- **Period Covered:** January, 2019
- **Channels:** E-mail, Texts, Verbatim feedback submitted through website
- Data covers complaints received by [Company XYZ] regarding products and service in six brands, henceforth referred to as [Brand 1] to [Brand 6].
- Products are referred to as [Product p.q] where p is the brand designation, q is product. Hence, [Product 1.2] is the second product mentioned for [Brand 1]

dbdayta@up.edu.ph

Consumer Data | First Looks

Processing the data introduced a number of challenges in cleaning.

Multilingual Texts

In general, complaints coming from the Philippines tend to be a mix of multiple languages and forms – English, Filipino, slang, regional languages.

dbdayta@up.edu.ph

Consumer Data | First Looks

Processing the data introduced a number of challenges in cleaning.

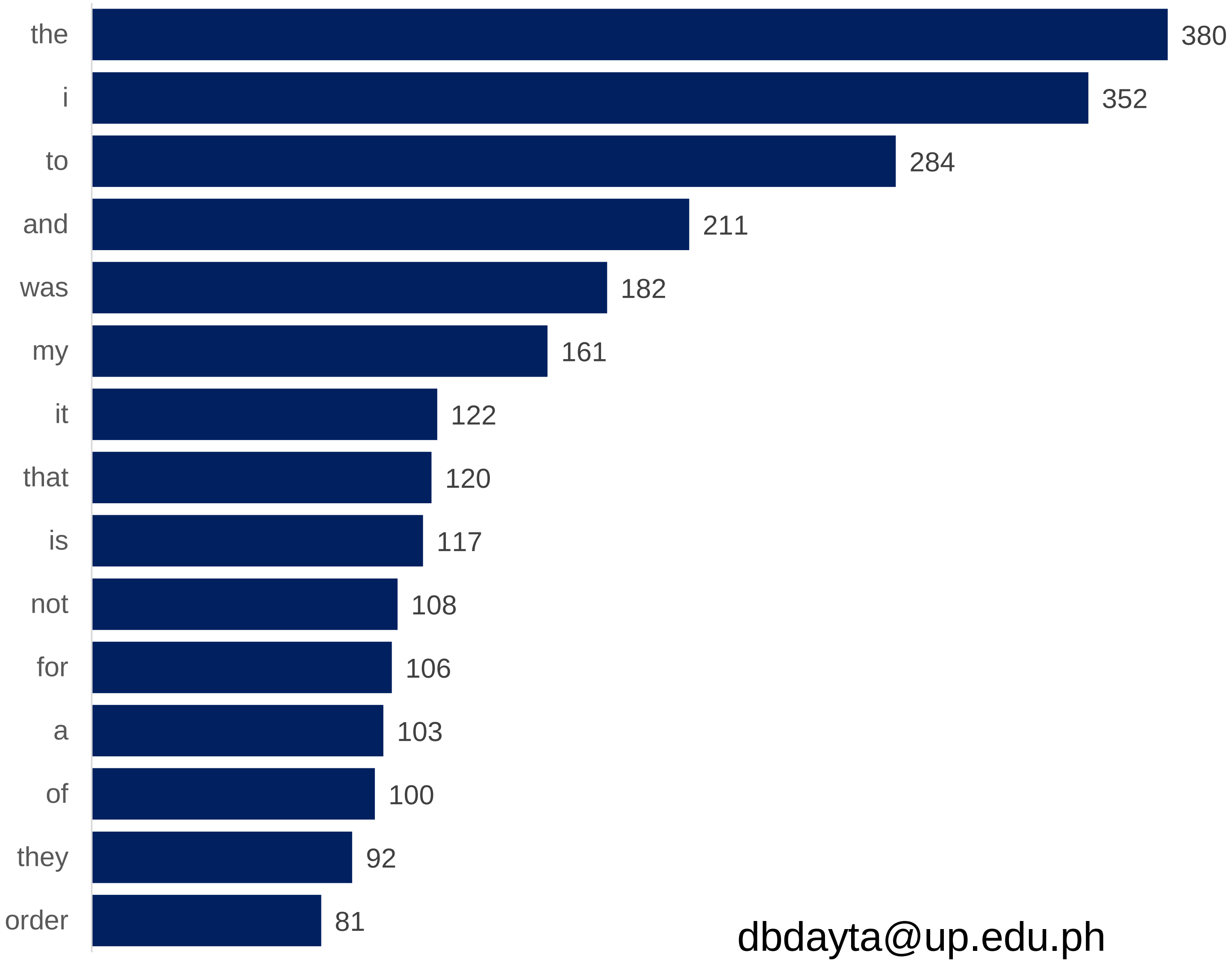
Lack of Formality in Writing

In general, Filipino writing exhibits a unique lack of formality. Words are contracted, rules of punctuation are ignored, grammar and basic construction are not observed.

dbdayta@up.edu.ph

Consumer Data | First Looks

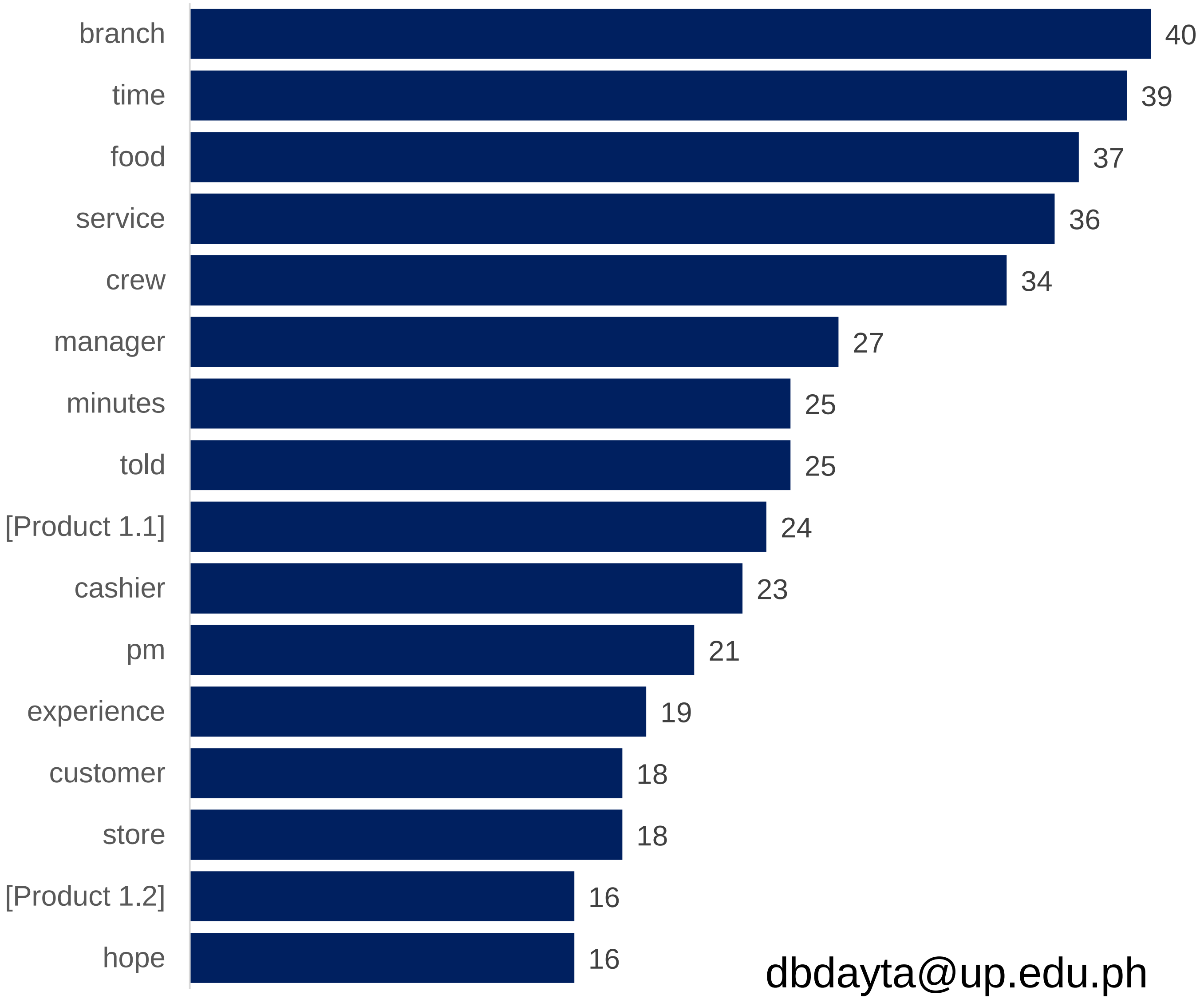
Most frequently used words. The corpus, in general, contains about 1,439 unique words.



dbdayta@up.edu.ph

Consumer Data | First Looks

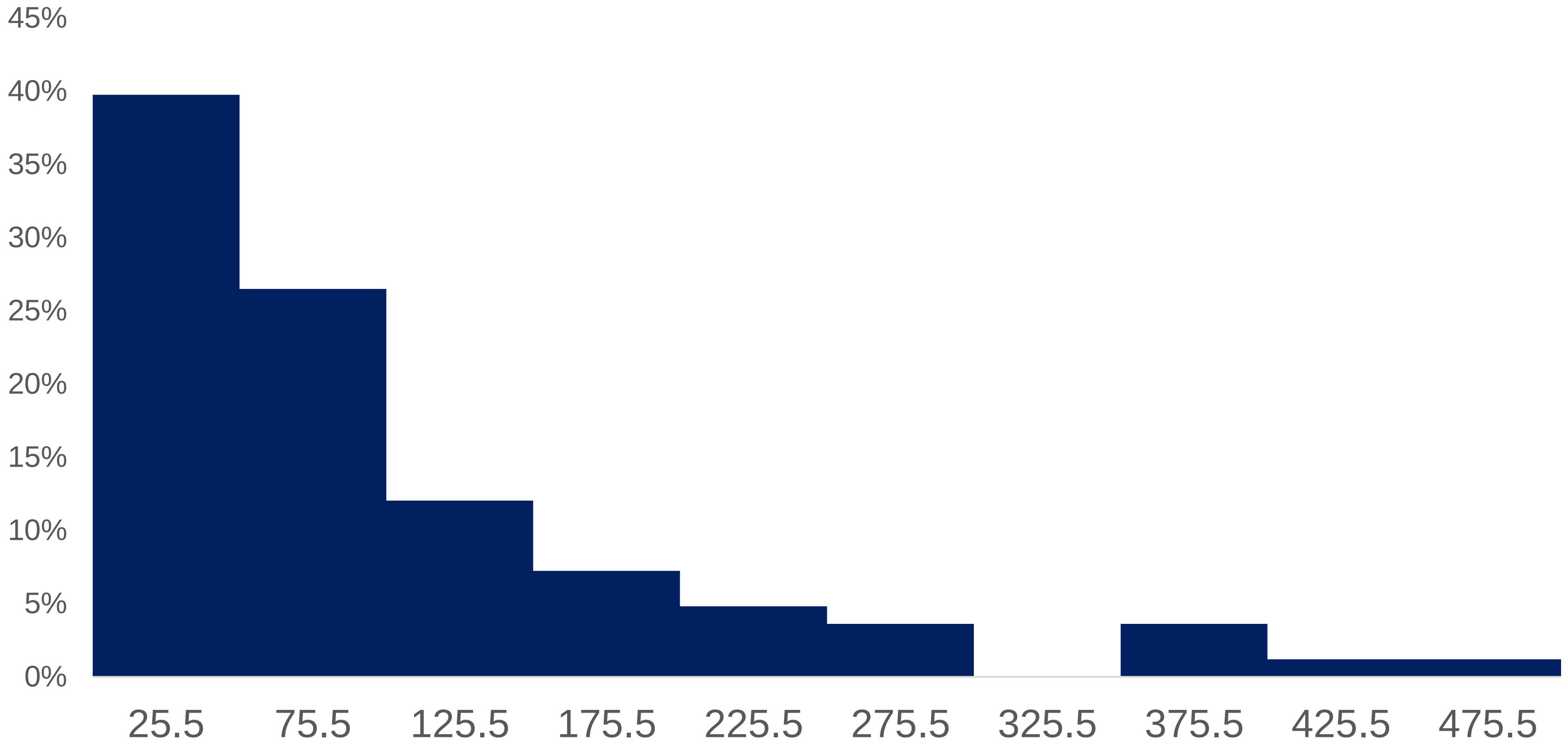
Most frequently used words, minus stop words. Stop words are functional words in the language that have no contextual meaning.



dbdayta@up.edu.ph

Consumer Data | First Looks

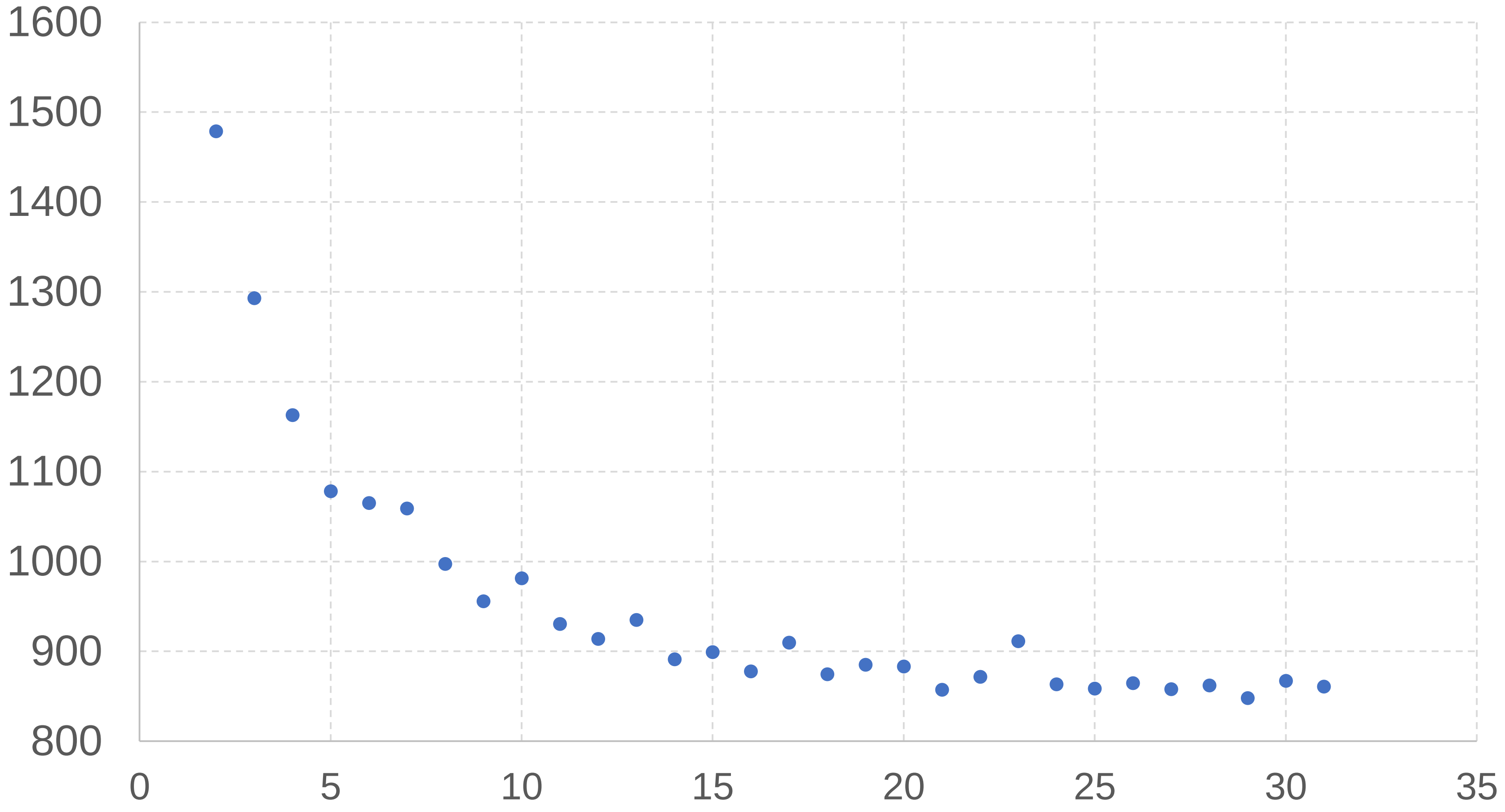
Complaints tend to be on the short side. On average, a complaint tends to be 104 words long, where the shortest complaint was 9 words only, and the longest at 457 words.



dbdayta@up.edu.ph

Results | Cross-Validation

Resulting inverse likelihoods (perplexity) for 2 to 31 topics. For the model, 15 topics were chosen.



dbdayta@up.edu.ph

Results | Topic Distribution 1

The first topic appears to concern, in order, the following words (with their corresponding importance value, β)

Top 10 Words

Time
Crew
[Product 1.1]
Fan
Food
[Product 2.1]

Disappointed
Talk
[Product 2.2]
Window
Replace
Forgot

Finally

dbdayta@up.edu.ph

Results | Topic Distribution 2

Meanwhile this topic appears to concern, in order, the following words (with their corresponding importance value, β)

Top 10 Words

Meal
[Brand 1]
[Product 1.2]
Hotline
[Product 1.3]
Square
[Product 1.4]
[Product 4.1]
Makati
Found

dbdayta@up.edu.ph

Results | Topic Distribution 3

Meanwhile this topic appears to concern, in order, the following words (with their corresponding importance value, β)

Top 10 Words

Branch	*Hot*	*Attention*
Service	*[Product 1.5]*	
City	*Share*	
Minutes	*January*	
Customers	*Bad*	

dbdayta@up.edu.ph

Results | Topic Distribution 4

Meanwhile this topic appears to concern, in order, the following words (with their corresponding importance value, β)

Top 10 Words

Eat	*sauce*
Feedback	*Customers*
Minutes	*Missing*
PM	*10*
Food	
Waiting	

dbdayta@up.edu.ph

Results | Topic Distribution 5

Meanwhile this topic appears to concern, in order, the following words (with their corresponding importance value, β)

Top 10 Words

Branch	*Line*
Crew	*Shocked*
Understand	*Incomplete*
People	*Evening*
Wait	*Counter*
[Product 1.4]	*bag*

Current Challenges

- So far, the method demonstrated is in keeping with the bag of words approach to text data. Language, however, is highly relative. Future implementation may consider, instead, n-grams (sequences of n consecutive words)
- The bag of word approach is naïve to linguistic phenomenon, e.g. polysemy. Words like “shot” which carry different contexts (photography, guns, sports, etc)
- Scaling the idea to industry level requires sophisticated data cleaning procedure for language problems discussed

dbdayta@up.edu.ph

Generalizations

- Companies and institutions can take advantage of emerging methods for text data to automate tedious data gathering processes related to customer feedback management
- Latent Dirichlet Allocation proves to be useful in automatically sifting through complaints according to subject
- Full implementation, however, requires some further thought into preprocessing, and scaling the method to industry size

dbdayta@up.edu.ph

End of Presentation

Thank You!

dbdayta@up.edu.ph