Topic Subject Creation Using Unsupervised Learning for Topic Modeling

Rashid Mehdiyev[1], Jean Nava, Karan Sodhi[1], Saurav Acharya[1] & Annie Ibrahim Rana[1]

[1] iON Regulatory Compliance, Underwriters Laboratories, LLC, Northbrook, IL, USA

Correspondence: Rashid Mehdiyev, iON Regulatory Compliance, Underwriters Laboratories, LLC, Northbrook, IL, USA.

Abstract

We address the problem of topic mining and labelling in the domain of retail customer communications to summarize the subject of customers inquiries. The performance of two popular topic mining algorithms - Non-Negative Matrix Factorization (NMF) and Latent Dirichlet Allocation (LDA) – were compared, and a novel method to assign topic subject labels to the customer inquiries in an automated way was proposed. Experiments using a retailer's call center data verify the efficacy and efficiency of the proposed topic labelling algorithm. Furthermore, the evaluation of results from both the algorithms seems to indicate the preference of using Non-Negative Matrix Factorization applied to short text data.

Keywords: topic mining, topic modeling, unsupervised machine learning, information retrieval

1. Introduction

The topic modeling domain of Natural Language Processing (NLP) has been quite popular in identifying the subject matter of a collection of documents, as well as the classification of documents (Berry, M. et al, 2009, Greene & Cross, 2016, Shah et al., 2018, Hingmire, S. et al.). The area of application for topic modeling has been rapidly expanding beyond NLP to computer vision (Shashua & Hazan, 2005, Chen et al, 2016), bioinformatics (Brunet et al., 2004, Liu et al., 2016, Mejía-Roa et al., 2008), recommender systems (Bao et al., 2014, Ju et al., 2015), astronomy (Berne et al., 2007, Zheng & Zhang, 2008, Saha, et al., 2015) and, many other areas.

The documents often need to be classified using tagging or labelling methods. However, the manual effort to perform these operations is too extensive, hence automating the tasks for topic mining and topic labelling is important. Traditionally, topic modeling and labeling techniques have been developed for long documents. Customer communications, on the other hand, are usually short conversations, most often noisy and imprecise, which makes the problem of topic identification challenging.

Topic models refer to the documents as a mixture of topics, and each topic consists of groups of related words, ranked by their relevance. Labelling in this context refers to finding one or a few single words or phrases that sufficiently describe the topic in question.

Automated topic labelling becomes an important matter in order to support users or customers in efficiently understanding and exploring document collections, as well as facilitating a reduction of manual efforts for the labelling process.

A large number of topic models and algorithms have been proposed to extract interesting topics in the form of multinomial distributions from the corpus in an unsupervised way.

The most popular ones are LDA (Latent Dirichlet Allocation), based on probabilistic modeling and Non-Negative Matrix Factorization (NMF), based on Linear Algebra.

Common features of these models are:

• The number of topics (k) needs to be provided as a parameter. Most of the algorithms cannot infer the number of topics in the document collection automatically.

• Both algorithms use Document-Word Matrix or Document-Term Matrix as input.

• Both of them output two matrices: Word-Topic Matrix and Topic-Document Matrix. The result of their multiplication should be as close as possible to the original document-word matrix.

LDA (Blei et al, 2003, Blei & Lafferty, 2006) uses Dirichlet priors for the word-topic and document-topic distributions. Each document may be viewed as a mixture of various topics where each document is considered to have a set of topics that are assigned to it via LDA. Topic distribution in LDA is assumed to have a sparse Dirichlet prior. LDA is a generative model that allows observations about data to be explained by unobserved latent variables that describe why some parts of the data are similar, or potentially belong to groups of similar topics. A topic in LDA is a multinomial distribution over the terms in the vocabulary of the corpus.

A different approach, such as NMF (Lee & Seung, 1999), has also been effective in discovering the underlying topics in text corpora (Greene & Cross, 2016). NMF is a group of algorithms in multivariate analysis and linear algebra, and in that way, it is essentially different from probabilistic methods used in LDA type of models. NMF is an unsupervised approach for reducing the dimensionality of non-negative matrices, which decompose the data into factors that are constrained so as to keep only non-negative values.

By modeling each object as the additive combination of a set of non-negative basis vectors, an interpretable clustering of the data can be, in principle, produced without requiring further post-processing. When applied to the textual data, these clusters can be interpreted as topics, where each document is viewed as the additive combination of several overlapping topics.

2. Method

The reasoning of this paper is to learn how effective these two very popular, albeit quite different topic modeling approaches could be applied to quite specific linguistic domain of relatively short-length customer communications with a specific vocabulary and terminology, as opposed to plain text corpora frequently tested in most topic modeling applications.

The overall approach used in this paper could be described by a number of processes, namely, data ingestion, data handling, processing, topic modeling, topic label generation, and analysis. The flow of these processes is presented in (Fig 1).

3. Data Preparation

The data studied consists of a corpus with 50,000 variable length (but mostly of few sentences long) text inquiries, originated from communications of commercial/retail company customers with the company's customer service personnel (also often known as log files). The subjects of the inquiries may vary greatly (thus, topic mining is needed), but, in our particular case of study, it refers mostly to television products.

The list of inquiries was ingested and extracted from the original customer service log files and pre-processed to cleanse the text. Standard text cleansing techniques like tokenization, case conversion, and stop words filtering have been applied to the original text. The pre-processing step also includes the removal of extremely short sentences, as well as, filtering of certain type of words, such as: named entities, personal names, overly frequent phrases and keywords specific to the nature of communications between customers and customer service department, by dropping the words like: 'customer', 'service', 'caller', days of the week, identical sentences, such as pre-prepared formal replies from the customer service.
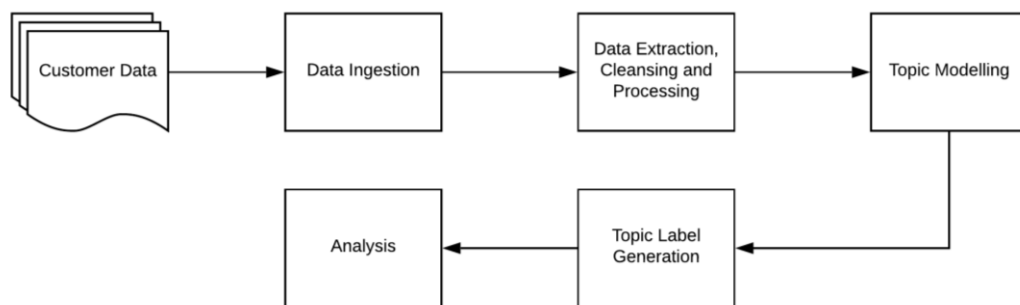

Figure 1. Schematic view of the end to end process flow in this paper

Additionally, duplicate words in the selected corpus, as artifacts of the process of splitting of textual input to sentences and further tokenization down to words, have been dropped as well. Typically, in the text processing domain, lemmatization and/or stemming of words are quite popular to remove tenses and plurals. The original text of inquiries (the log files) also contain a fair amount of misspelled words and typographic mistakes. No systematic attempts were used to correct those typos, as it might hinder the idea of automation of label generation. During text-processing, we have tested two options, with word lemmatization, and without it.

For the two topic modeling approaches studied in this paper, the same text pre-processing of raw data was conducted and then the resulting text was fed as input to each model in order to perform topic modeling.

Pre-processed data resulted in 40,000 observations (we call them snippets, for the rest of the paper) for each model, for testing. The number of topics as 40, the number pre-defined a-priori, has been used as a parameter for both models to be compared. No attempt was made to use topic coherence study or similar methods to optimize the number of topics automatically from the bulk of data. This will be studied in our next paper.

4. Topic Modeling

In the bag-of-words model, each document is represented by a vector in a m-dimensional coordinate space, where m is number of unique terms across all documents. This set of terms build the corpus vocabulary. Since each document can be represented as a term vector, we can accumulate these vectors to create a full document-term matrix. We can create this matrix from a list of document strings (an inquiry, in our case). From our data, we have created (39697×434) TF-IDF-normalized document-term matrix, let's call it matrix V. The usefulness of the document-term matrix is justified by giving more weight to the more "important" terms. The most common normalization is widely known as Term Frequency–Inverse Document Frequency (TF-IDF). With scikit-learn library [scikit], by using the TfidfVectorizer method, we can generate a TF-IDF weighted document-term matrix.

In the mathematical discipline of linear algebra, a matrix decomposition or matrix factorization is a factorization of a matrix into a product of matrices. By applying matrix decomposition to document-term matrix V, NMF produces two factor matrices as its output: W and H. In a formal way, V matrix decomposition could be presented as $V_{(n \times m)} \approx W_{(n \times k)} \times H_{(k \times m)}$. The W matrix contains the document membership weights relative to each of the k topics. Each row in the W matrix corresponds to a single document, and each column correspond to a topic. The H matrix contains the term weights relative to each of the k topics. In this case, each row corresponds to a topic, and each column corresponds to a unique term in the corpus vocabulary.

The top ranked terms (or descriptors) from the H matrix for each topic can give an insight into the content of the topic.

On the other hand, the LDA model can only use raw term counts/frequencies because LDA is a probabilistic model, that uses probabilities of words across the corpus. Thus, as opposed to NMF and Tfidf vectorizer, scikit-learn CountVectorizer method has been used to count LDA originated terms. Total number of 445 terms were found in 39,697 input documents, if lemmatization was applied. With no lemmatization the statistics were 518 terms in 39,474 documents, respectively. Final decision was taken to proceed with lemmatization, as, it reduces the number of terms originated from closely related words.

In order to compare two models, we have constructed (similarly to NMF), W and H matrices, but based on LDA model output.

An important step in topic modeling is to produce a set of terms (also known as descriptors) which characterize topics discovered in the modeling. The list of terms for each of the topic (limited to 10 topics), as found by each respective model are presented below. Table 1 corresponds to the results of NMF, while Table 2 corresponds to the results of LDA.

The list of terms, provided by NMF model are ranked for each topic by the term weight, obtained from the matrix H of the NMF model. So, the first term, in the top row of each topic column in Table 1, could be considered as an initial candidate for a corresponding topic label. One can notice also, that for NMF, the highest weighted term typically has a rather close semantic relationship with the rest of the terms of the same topic.

In the case of LDA, as can be observed in Table 2, due to its probabilistic approach, the model tends to over-represent the most probable term across many topics (e.g. consider the term "tv" which is omnipresent across the list of terms shown).

The graphical side-by-side examples of distributions of topical terms/descriptors obtained from NMF and LDA, are presented below for 5 potentially matching topics.

For the sake of a graphical comparison between the two models, the term weights for NMF were normalized to the highest term weight for each topical term distribution, while the normalization in the LDA's case was done to scale the terms.

Table 1. Ten topics with topical terms obtained from NMF

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
turn	status	power	screen	sound
keep	followup	cycling	crack	pop
tv	inquiry	cord	half	cut
automatically	inquire	cycle	white	hear
report	verify	button	blue	speaker
intermittently	chat	lead	flash	display
longer	live	reporting	dark	video
anymore	phonecell	television	flicker	bar
time	promise	report	spot	problem
onoff	see	anymore	damage	click
Topic 6	**Topic 7**	**Topic 8**	**Topic 9**	**Topic10**
update	ticket	line	picture	remote
software	cancel	vertical	audio	control
firmware	create	horizontal	flicker	pair
usb	recreate	bottom	show	button
account	review	top	see	respond
access	creation	screen	dark	chat
date	open	color	sent	live
try	chat	middle	reception	smart
purchase	reject	green	send	defective
drive	live	red	dim	battery

Table 2. Ten topics with topical terms obtained from LDA

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
turn	status	tv	screen	connect
tv	check	power	tv	tv
know	information	cycling	side	box
bos	processing	visible	left	internet
pending	ticket	appear	dark	time
randomly	step	month	right	try
adjust	rma	told	set	center
attach	distort	flicker	half	replace
direct	quick	unit	damage	talk
status	add	show	darker	wifi
Topic 6	**Topic 7**	**Topic 8**	**Topic 9**	**Topic 10**
update	ticket	line	picture	remote
ask	process	vertical	sound	work
connection	inform	display	tv	control
pick	open	spot	white	tv
transaction	locate	middle	intermittently	replacement
software	tv	horizontal	bar	chat
process	credit	inch	sometimes	live
wireless	software	green	bought	properly
firmware	follow	multiple	distort	help
version	refund	gray	lose	rep
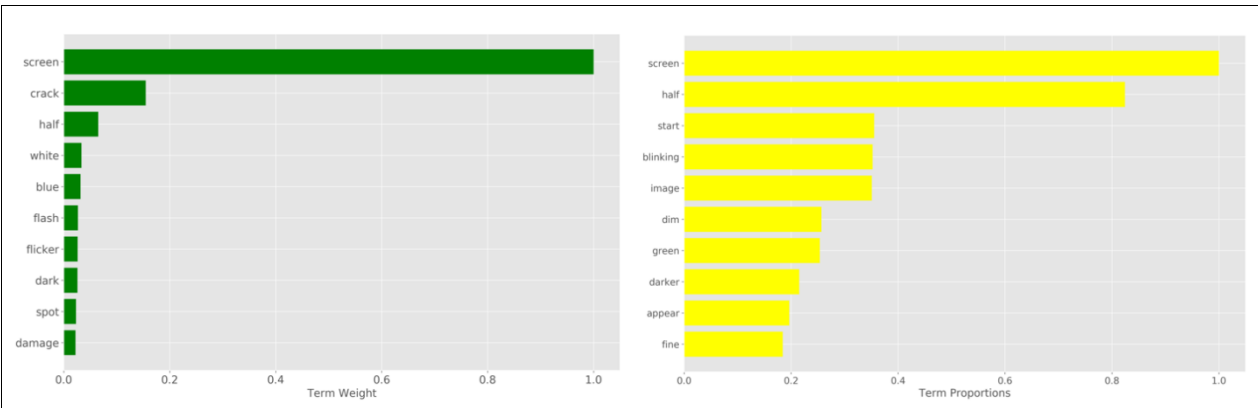
"Screen" topic terms in NMF (on left) and in LDA (on right)
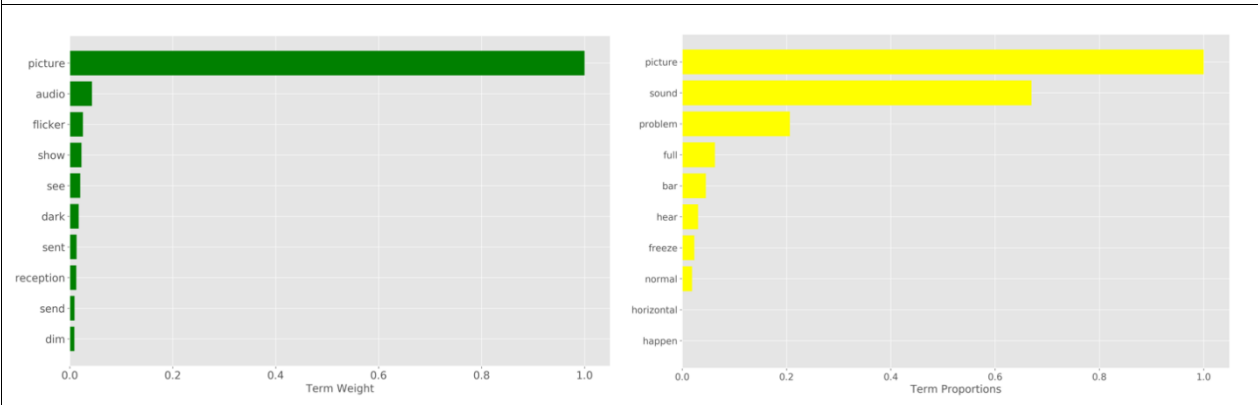
"Picture" topic terms in NMF (on left) and LDA (on right)

"Remote" topic terms in NMF (on left) and LDA (on right)

"Line" topic terms in NMF (on left) and LDA (on right)
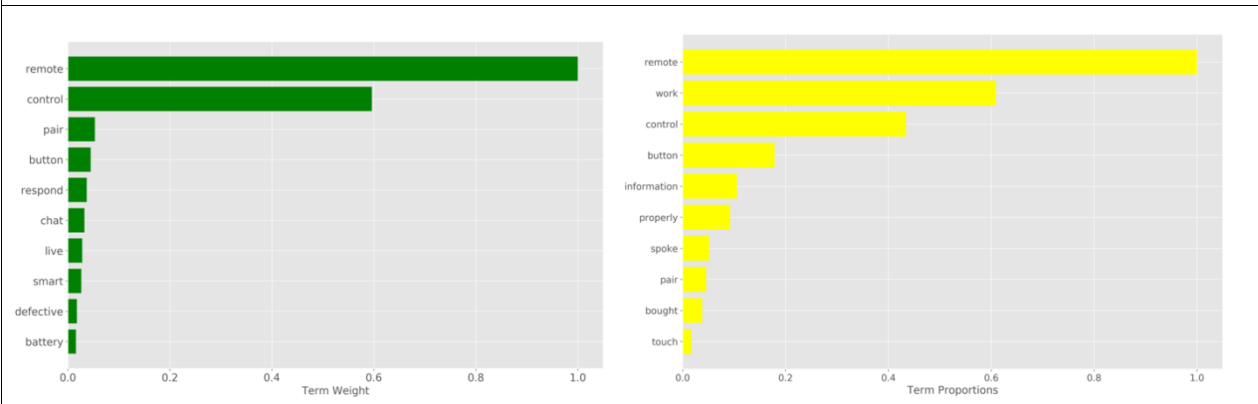
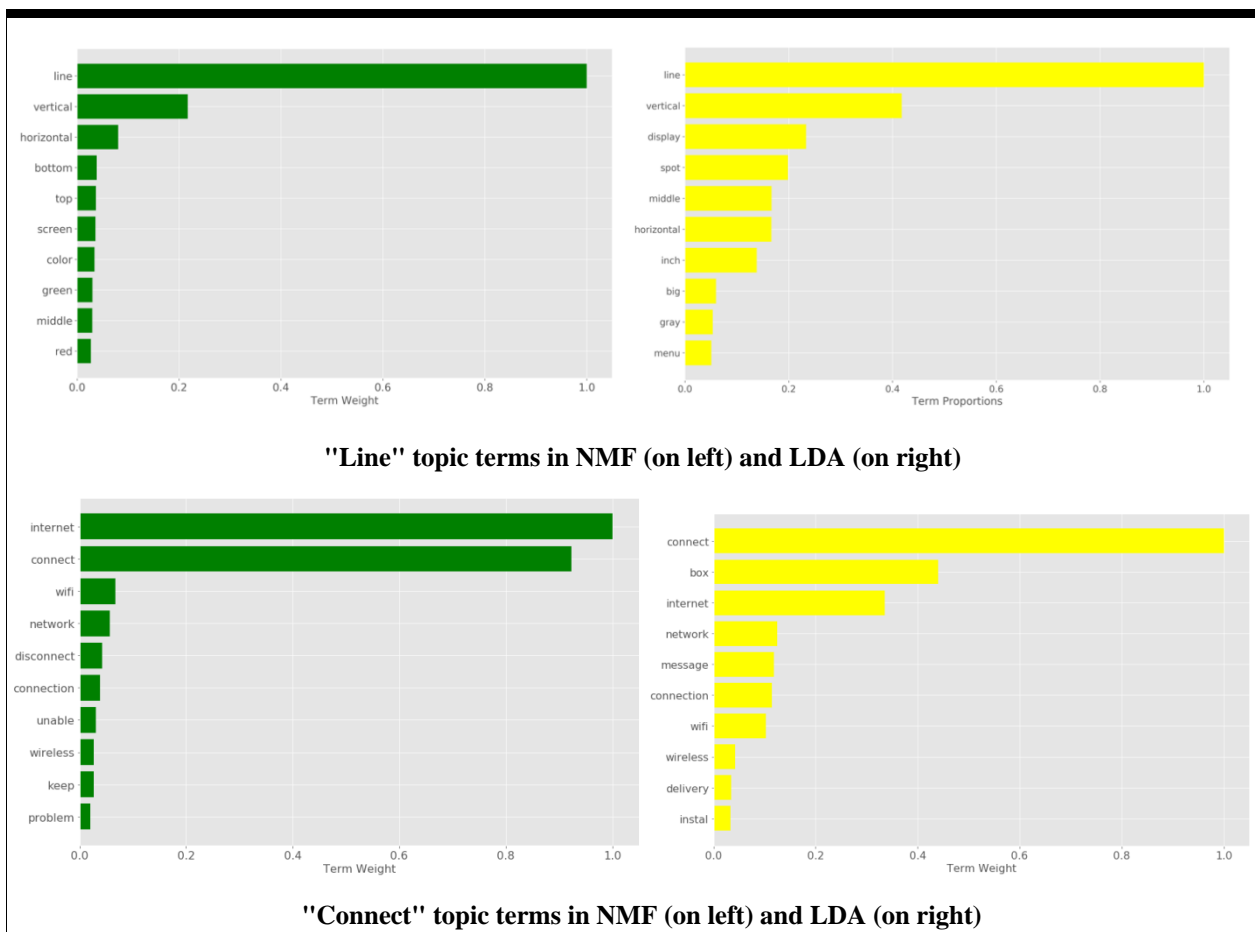**"Connect" topic terms in NMF (on left) and LDA (on right)**

Figure 2. Comparison of top 10 term distributions in NMF (green) and LDA (yellow) for 5 selected topics

distributions to one. This graphical representation helps to visualize the striking difference in topical terms resulting from two models studied. The term distribution in NMF, as a consequence of using TF-IDF method, tends to be mostly dominated by a term with the highest weight. For LDA, the term distributions are much wider presumably due to the fact that a simple counting of words is less efficient in picking up most representative term/word for each topic. Thus, in our opinion, the counting of words is prone to picking up the terms semantically but not always close to the rest of the terms in each topic category. As an example, term representations in the LDA case often look like a boiler plate, showing, for this particular television related data, all kinds of TV related terms, but lacks the terms which would identify the label more or less unambiguously. It is lesser of a feature for NMF terms distributions.

5. Topic Labelling

Now we are in the position to generate a label for each document (inquiry) using the set of terms, or descriptors, obtained from the previous step.

The idea behind finding the top document for label generation is that, within the most "representative" document there is a text fragment that could contain a coherent label. This is a label that is as grammatically correct as possible (not always easy to do, taking into account that we decided to use lemmatization as a part of text preprocessing) and would be easily comprehended by humans. It is a challenging task to create labels as close as possible to human assigned labels, while being as representative and simple as possible.

For this attempt to come to a reasonable label that could be understood by humans, and help them grab a decent idea on the nature of a consumer inquiry, we select topics labels with highest ranked terms with the following steps:

1. Further lemmatization to construct and/or select only nouns from the obtained set of descriptors/terms.

2. The cosine similarity between the terms for each topic and all snippets (original inquiries) returns a list of snippets with a highest score.

3. Sentences/snippets have been ranked with LexRank algorithm (Erkan & Radev, 2004).

4. Using sentence similarity, the results for three top ranked sentences were kept as most relevant.

5. Only one candidate for a topic label, as the most similar to the majority of snippets selected above, was chosen as an ultimate label.

The same algorithms has been applied to both set of topical terms or descriptors derived from each model.

In the following two tables, for each model respectively, we present the 12 most representative cases (of a predefined number of 40 topics for each model). The order of topics for NMF model (Table 3) generally follows the ranking of terms by their assigned weights in the model. The 12 examples for LDA (Table 4) below were chosen by their similarity to NMF examples.

The results of the label generation show quite satisfactory matching patterns between original inquiries and generated labels. It should be noted that the application of lemmatization resulted in partial distortion of the final label grammatically that makes them a bit robotic. However, in the presented examples for the case of NMF, almost 90% of the topical terms are covered by generated labels quite well: out of 12 topics shown, the label for Topic 9 (Table 3) is probably a little bit vague.

For LDA generated labels (shown in Table 4), the mismatch between top snippets, descriptors and resulting generated labels seems to be more visible. Obtained labels for Topics 2 and 3 seems to be drawn from overlapping top snippets, For Topic 5, the list of terms leaves little choice to label between "line on screen" and "screen cracked". Similarly, for Topic 10, it is a difficult choice between "melted screen" and "spot on tv screen". It appears that the labels generated and based on LDA terms are slightly less accurate than in the NMF case.

The last column in Tables 3 and 4 shows a count of how many times a generated label was able to find a pattern in 1000 snippets used to validate the method. The cosine similarity tool has been used to compare the labels and the snippets/inquiries.

The comparison of hits (counts) show that the labels generated with the NMF model are more frequently able to find a match between snippets. A possible reason for the better performance of NMF is that the TF-IDF method, that is exploited in NMF is more adequate for the topical term selection than the term selection by word frequencies/proportions used in the LDA model. Taking into account the multitude of attributes with no particularly strong predictors in the bulk of the textual data (short communication logs) used in this study, the weighting of terms by importance in NMF is shown to work better in representing patterns and topics.

Table 3. NMF label generation results (12 topics shown)

Topic	Descriptors	Top snippets	Label	Hits
1	turn, keep	tv turn not, tv not turn on, tv turn off itself	tv turn not	747
2	status, inquiry, verify	exchange status ticket, service ticket status, status of ticket	exchange status ticket	248
3	follow, reception, tkt, order, pending	ticket follow up, follow on ticket, follow up ticket	ticket follow up	536
4	power, cycling, cord	tv power cycling, power cycling constantly, power cycling itself	tv power cycling	403
5	line, bottom, top	line on the screen, line across the bottom of screen, tv have line	line on the screen	534
6	work, button, source	button not work, source not work, other button not work	button not work	472
7	picture, audio, flicker, show	picture be flicker, no picture audio, tv show no picture	tv show no picture	225
8	update, software, usb, firmware, usb, account	software update request, update the software via usb, about the software update	about the software update	53
9	know, warranty, happen, sent, process	because the want to know process, want know about the exchange process, want to know the warranty coverage	because want to know process	738
10	side, left	uneven brightness left side be darker, dark spot on the whole left hand side, the left hand side of screen	the left hand side of screen be dark	341

		be dark		
11	check, dlr, tech, sent, technician	check ticket status, check status of, check exchange status	check ticket status	382
12	hdmi, port, recognize, device	hdmi port issue, hdmi port not work, no hdmi port work	no hdmi port work	318

Table 4. LDA label generation results (12 topics shown)

Topic	Descriptors	Top snippets	Label	Hits
1	tv ,turn, show, cancel	tv turn off, tv turn itself, tv turn not	tv turn off	499
2	status, know, tv, approve	tv exchange status, tv exch status, ask about tv status	tv exchange status	238
3	follow, tv, company, trucking, reception	tv follow up status, tv exchange follow up, follow up call tv service	tv follow up status	227
4	told, week, cycling, buy	tv power cycling, unit be power cycling, tv be power cycling	tv be power cycling	116
5	line, screen, tv, crack	tv screen crack, tv screen have line, tv line on screen	tv line on screen	542
6	button, support, tv	voice control button, contact live support tv screen get damage, home button not work on the tv	home button not work on the tv	128
7	picture, audio, happen, hear	get audio but no picture, distort line picture, get no picture but have audio	get audio but no picture	190
8	process, software, give, update, version	software update request, gpca do not process for software update, say they already update the software version	software update request	6
9	status, know, tv, approve	tv exchange status, tv exch status, ask about tv status	tv exchange status	238
10	screen, tv, side, left	black screen on left side of tv, tv screen be melt, spot on tv screen	spot on tv screen	61
11	ticket, check, status, provide, info	check ticket status, check ticket info, check the status ticket	check ticket status	406
12	hdmi, port, tv, device, television	hdmi port issue, tv hdmi port not recognize the, tv hdmi port be not work	hdmi port be not work	413

6. Conclusions and Future Work

Non-Negative Matrix Factorization (NMF) and Latent Dirichlet Allocation (LDA) algorithms are used in this study for topic mining and topic labelling, applied to customer textual communications to characterize the subject of customers inquiries. A method to assign generated topic labels has been proposed in attempt to make it as less human assisted as possible. The comparison of both algorithms seems to indicate the preference of using Non-Negative Matrix Factorization for the particular short text data. In the future, we plan to extend the work to research evolution of the topics over time.

References

Bao, Y., Fang, H., & Zhang, J. (2014). *TopicMF: Simultaneously Exploiting Ratings and Reviews for Recommendation.* Artificial Intelligence, Twenty-Eighth AAAI Conference, Quebec City, QC, Canada.

Berne, O. et al. (2007). Analysis of the Emission of Very Small Dust Particles from Spitzer Spectro-imagery Data Using Blind Signal Separation Methods. *Astronomy & Astrophysics, 469*(2). https://doi.org/10.1051/0004-6361:20066282

Berry, M., Gillis, N., & Glineur, T. (2009). *Document Classification Using Nonnegative Matrix Factorization and Underapproximation.* Circuits and Systems, IEEE International Symposium, IEEE. https://doi.org/10.1109/ISCAS.2009.5118379

Blei, D., & J. Lafferty, J. (2006). *Dynamic Topic modeling.* Machine Learning, 23rd International Conference, Pittsburgh, PA. https://doi.org/10.1145/1143844.1143859

Blei, D., Ng, A., & Jordan, M. (2003). Latent Dirichlet allocation. *The Journal of Machine Learning Research, 3.*

Brunet, J. P. et al. (2004). Metagenes and Molecular Pattern Discovery Using Matrix Factorization. *PNAS, 101*(12). https://doi.org/10.1073/pnas.0308531101

Chen, C., Zare, A., & Cobb, J. T. (2016). *Partial Membership Latent Dirichlet Allocation for Image*

Segmentation. Pattern Recognition Workshops (ICPR), IEEE International Conference, IEEE. https://doi.org/10.1109/ICPR.2016.7899990

Erkan, G., & Radev, D. (2004). LexRank: Graph-based Lexical Centrality as Salience Text Summarization. *Journal of Artificial Intelligence Research, 22,* 457-479. https://doi.org/10.1613/jair.1523

Greene, D., & Cross, J. (2016). *Exploring the Political Agenda of the European Parliament Using a Dynamic Topic Modeling Approach.* arXiv preprint arXiv:1607.03055.

Hingmire, S. et al. (2013). Document Classification by Topic Labeling. *Research and Development in Information Retrieval (ACM SIGIR), 20.* https://doi.org/10.1145/2484028.2484140

Ju, B. et al. (2015). Using Dynamic Multi-Task Non-Negative Matrix Factorization to Detect the Evolution of User Preferences in Collaborative Filtering. *PloS One, 10*(9), e0138279. https://doi.org/10.1371/journal.pone.0138279

Lee, D., & Seung, H. (1999). Learning the Parts of Objects by Non-negative Matrix Factorization. *Nature, 401,* 788-791. https://doi.org/10.1038/44565

Liu, L. et al. (2016). An Overview of Topic Modeling and its Current Applications in Bioinformatics. *SpringerPlus, 5*(1), 1608. https://doi.org/10.1038/44565

Mejía-Roa, E. et al. (2008). bioNMF: a web-based tool for nonnegative matrix factorization in biology. *Nucleic Acids Research, 36.* https://doi.org/10.1093/nar/gkn335

Saha, S. et al. (2015). *ASTROMLSKIT: A New Statistical Machine Learning Toolkit: A Platform for Data Analytics in Astronomy.* arXiv preprint arXiv:1504.07865.

Shah, A. M. et al. (2018). Use of Sentiment Mining and Online NMF for Topic Modeling Through the Analysis of Patients Online Unstructured Comments. *ICSH 2008 Smart Health,* 191-203. https://doi.org/10.1007/978-3-030-03649-2_19

Shashua, A., & Hazan, T. (2005). *Non-negative Tensor Factorization with Applications to Statistics and Computer Vision.* Machine Learning (ICML) '05, 22nd international conference on, Bonn, Germany. https://doi.org/10.1145/1102351.1102451

Soukup, D., & Bajla, I. (2008). Robust Object Recognition under Partial Occlusions Using NMF. *Computational Intelligence and Neuroscience, 857453*. https://doi.org/10.1155/2008/857453

Zheng, H., & Zhang, Y. (2008). Feature Selection for High Dimensional Data in Astronomy. *Advances in Space Research, 41,* 1960-1964. https://doi.org/10.1016/j.asr.2007.08.033

Copyrights