Topic Modelling on Pharmaceutical Incident Data

Deepu Dileep, Soumya Rudraraju, and V. V. HaraGopal

***Abstract* — Focus of the current study is to explore and analyse textual data in the form of incidents in pharmaceutical industry using topic modelling. Topic modelling applied in the current study is based on Latent Dirichlet Allocation. The proposed model is applied on a corpus containing 190 incidents to retrieve key words with highest probability of occurrence. It is used to form informative topics related to incidents.**

***Index Terms* — Coherence Score, Incidents, Latent Dirichlet Allocation (LDA), Textual Mining, Topic Modelling.**

I. Introduction

Pharmaceutical industry is a reservoir for abundant data with huge potential for analytical insights. Advanced techniques are to be employed in harnessing this potential of the data. Textual mining and data analytics are employed in the current study to extract essential information from text data. Topic modelling have proved to be essential in extracting key information from a set of documents. Among different Topic modelling techniques, Latent Dirichlet Allocation is proved to be the most reliable technique for the purpose of keyword extraction [1]-[3].

Incidents are unplanned events that exceed limits, specifications or expectations. Incidents may or may not have an impact on quality. So, it is necessary to conduct an investigation on the incidents to derive its importance. This can lead to the identification of its root causes and the effective Corrective and Preventive Action (CAPA) to prevent. In terms of CAPA's, both a quality-impacting incident or a deviation, in most if not all cases, require an investigation that gets to the root causes and comes up with effective CAPA's that prevent occurrence and recurrence.

Topic modelling based on LDA has been researched and widely applied for various datasets. LDA is used to extract key words and to identify important topics from Prime Minister of India, Narendra Modi's Mann Ki Baat [4]. Difference in Topics are identified by calculating Topic coherence as it provides higher correlation with human ranking compared to loglikelihood [5], [6].

The current paper is an attempt to apply an unsupervised machine learning model text data. The method used is Topic modelling based on LDA. It is applied on 190 incident text data to extract the most essential key words which help in future incidences. LDA creates a set of topics with the combination of words from the corpus. These words are distributed in a topic based on their probabilities. LDA clusters each text into different topics based on the probability of occurrence. Our study highlights the prominent keywords among the incidents based on highly probable topic associated with each incident. This research is the first to explore and analyse pharmaceutical incident data using LDA to extract key words and form informative topics.

II. Materials and Methods

A. Dataset

The data used for the current study is pharmaceutical text data in the form of incidents. These incidents occurred at a Drug Research Solutions firm. A total of 190 incidents from 10 different departments were used to purse the analysis.

B. Data Pre-processing

Data Pre-processing is a vital part in any text modelling. It's performed to remove the noises present in the raw data. The process of Pre-processing involves conversion to lower cases then removal of punctuations, symbols, and extra spaces. The text data is then tokenized. Stop words are removed from the tokenized data. The remaining words are fed to a technique named lemmatization [7] to convert the words into their base or dictionary form of a word. This process is visualised in Fig. 1.

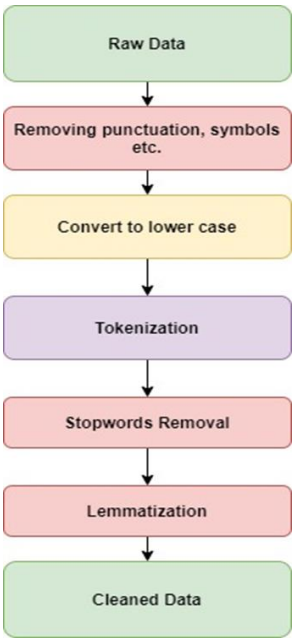

Fig 1. Pre-processing flow chart.

Published on July 30, 2021.
Deepu Dileep, Aizant Global Analytics Pvt Ltd, India.
(e-mail: deepu.dileep@aizantit.com)
Soumya Rudraraju, Aizant Global Analytics Pvt Ltd, India.
(e-mail: rsoumyaraju@gmail.com)
V. V. HaraGopal, Aizant Global Analytics Pvt Ltd, India.
e-mail: haragopal.vajjha@aizantit.com

C. Topic Modelling using LDA

Topic Modelling is an unsupervised technique for natural language processing. It provides a probabilistic framework for organizing, analysing, and summarizing text data. Topic modelling creates each topic as a probabilistic distribution of words from the input documents. Among various topic modelling techniques Latent Dirichlet Allocation is found to be most efficient and popular. LDA assumes each document to be a mixture of topics and each topic to be a mixture of words. The process of LDA is summarised visually in Fig. 2.

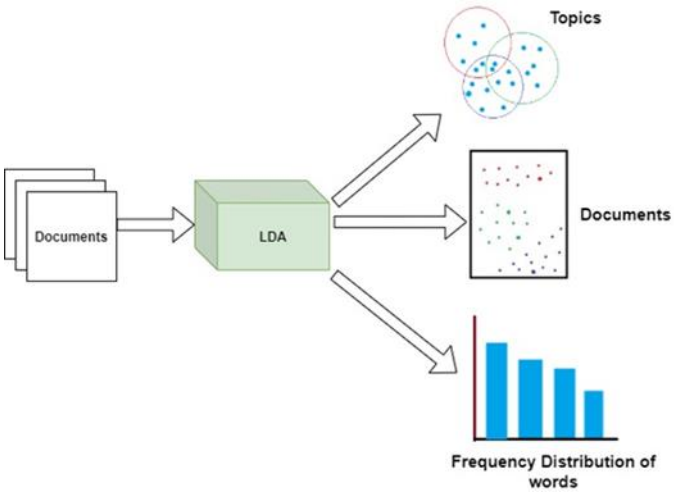

Fig 2. LDA Flow chart.

The process in LDA contains two Dirichlet distributions and two multinomial distributions. These distributions are used to obtain the probability of a document. Probability of a document is given in Fig. 3.

$$P(W,Z,\theta,\varphi,\alpha,\beta) = \cdot \prod_j^M P(\theta_j;\alpha) \prod_i^K P(\varphi_i;\beta) \prod_t^N P(Z_{j,t}|\theta_j)\ P(W_{i,t}|\varphi Z_{j,t}) \quad (1)$$

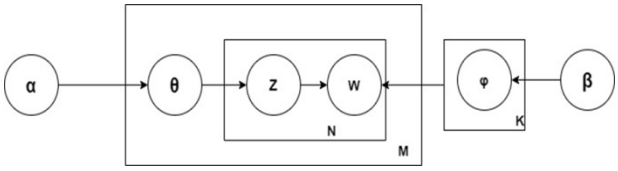

Fig 3. LDA representation.

LDA represents a corpus of probability of M documents. It uses the Dirichlet distribution with parameter α to distribute documents between topics. This in turn is clubbed with multinomial distribution (θ), this represents the probability distribution of a topic chosen from a document. The second Dirichlet distribution with parameter β represents the distribution of topics between words. A multinomial distribution (ϕ) to represent the probability of a word from given topic.

Here K represents the number of topics, N the number of words, Z represents topics and W represents words in documents.

Topic coherence is used to select the best model parameters including the number of topics. Topics are considered to be coherent if the words are related. Coherence is based on C_v coherence score, it segments data into word pairs and calculates word pair probabilities. It quantifies a word set with another and aggregates these individual scores by computing the mean [8].

III. Experimental Results

The experiment was conducted on incident data using Python 3. NLTK, Genism and Regex libraries are employed to achieve the required results. The process included text pre-processing followed by application of LDA. Optimal number of topics was chosen based on the coherence score and the minimum number of common words between topics and the distribution is summarised in Table I.

TABLE I: Representation of Different Number of Topics along their Coherence Score and Total common Words

No of Topics	Coherence Score	Common words between Topics
6	0.5003	52
7	0.5038	63
8	0.5091	58
9	0.5153	77
10	0.5175	59
11	0.5225	73
12	0.529	60
13	0.5306	80
14	0.5372	87

From the table it is evident that as the number of topics increases the coherence score tends to increase. But the common words between topics also tend to increase with it. Considering a suitable coherence score and minimum number of common words optimal number of topics were found to be best when number of topics equals 12.

A parametric study is conducted on model with 12 topics and it was found that the Dirichlet parameter alpha holds a huge effect on the coherence score and common words between topics. It is visually represented in Fig. 4 and 5.

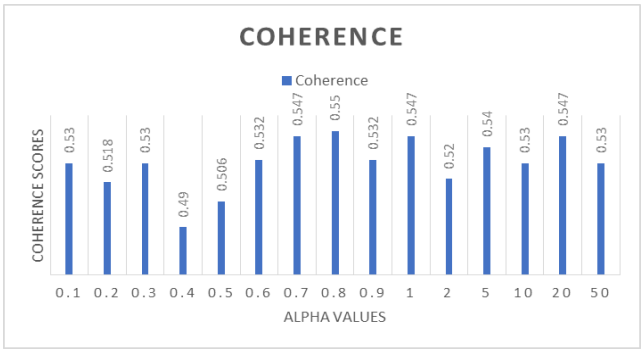

Fig 4. Coherence score vs alpha values.

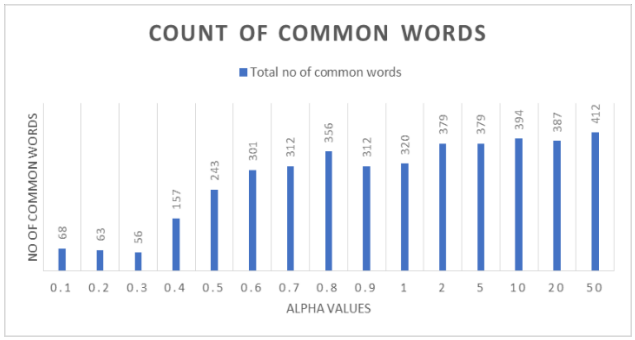

Fig. 5. Total no of common words vs Alpha.

The parametric study shows that higher the value of alpha Parameter higher the number of common words is found between the 12 topics. It is desired to have an alpha with least number of common words and higher coherence score. Alpha as 0.3 is identified as the optimal parameter as it satisfies both the condition.

LDA model is built using the optimal parameters and applied on to the corpus. The top 20 high probability words for each topic are retrieved and are shown in Tables II and III.

Probability distribution of each topic over the incident data is plotted in Fig. 6. The topics with maximum probability corresponding to each incident can be observed as peaks in the graph. The peak for each incident is identified and represented as the number of incidents represented by each topic. Frequency of occurrence of different topics are plotted in Fig. 7.

TABLE II: TOP 20 WORDS IN TOPICS 1 TO 6

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6
number	drying	report	tablet	form	coating
sample	monthly	calibration	production	lab	done
generation	production	found	blend	steam	method
cabinet	common	stability	filter	located	sequence
label	supervisor	hplc	qa	immediately	chromatogram
instrument	perform	daily	layer	cycle	tablet
using	blend	missing	personnel	withdrawn	time
serial	granulation	protocol	observed	month	packaging
multiple	dried	detail	pack	rpm	bottle
hplc	capsule	sample	warehouse	report	preparation
fixed	chamber	analytical	executive	following	taken
one	unit	quantity	archival	full	analysis
incharge	operation	validation	execution	external	blister
system	lot	timeline	packaging	finished	sterilizer
dissolution	calibration	lab	monitoring	temp	given
generated	chlorpromazine	tablet	black	working	limit
lopinavir	loss	software	quarantine	swab	started
ool	coated	methocarbamol	code	protocol	used
std	due	released	feeding	criterion	sign
register	maleate	slip	residue	procedure	maintenance

TABLE III: TOP 20 WORDS IN TOPICS 7 TO 12

Topic 7	Topic 8	Topic 9	Topic 10	Topic 11	Topic 12
temperature	raw	qualification	compression	meq	water
range	data	performed	batch	cardisar	audit
incubator	page	le	weight	missed	area
standard	calculation	microbiology	coated	potassium	vendor
within	logbook	temperature	yield	usp	observation
recorded	potency	stage	instead	date	laboratory
microbiology	inward	yield	film	citrate	personnel
review	review	lab	record	standard	limit
located	date	compression	ptsti	due	purified
revised	revisied	conductivity	release	valta	client
balance	manditory	supplementary	tablet	failure	shown
time	lnb	expiry	observed	ups	test
went	verifcation	date	limit	analysis	found
observed	material	micropipettes	digit	wrong	granule
period	arno	incubator	printed	coding	note
test	tablet	channel	ppb	micropipette	instead
listed	chromatograme	stability	process	amcc	nmt
attachment	missing	observed	methocarbamol	pattern	book
material	power	range	block	quality	bioburden
laboratory	happened	allergex	extended	blister	april

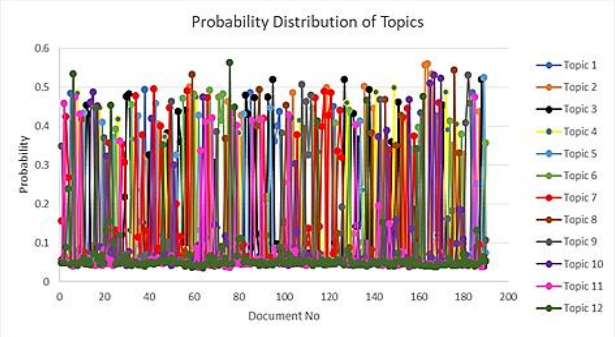

Fig. 6. Probability Plot of Topics Extracted for Incidents.

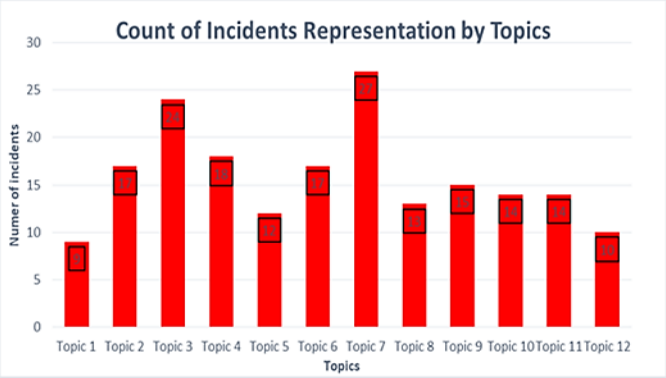

Fig. 7. Topic Frequency Distribution for Incident.

From Fig. 7 it can be observed that topic 7 represents the maximum number of incidents followed by topic 3.

For analyzing most popular words in the incident corpus a sparse matrix is used to represent the incidents versus words frequency table, most popular words among the corpus are represented in Fig. 8 with their frequency of occurrence.

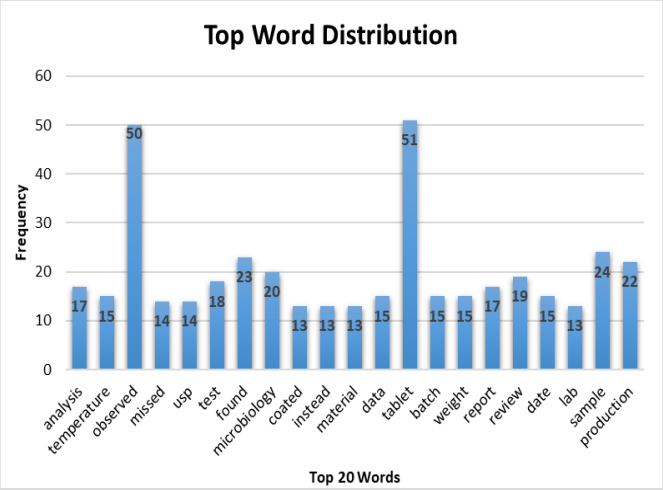

Fig. 8. Word Distribution.

IV. Conclusion

In this paper we investigated on employing Topic modelling using LDA on Incident text data to extract prominent key words. Based on the Analysis the corpus of 190 incidents was distributed between various topics based on the probability. A total of 12 topics were derived from the corpus based on coherence score and minimum number of common words between topics. The model can assign the incidents into different topics based on the highest probability. For a new incident the model can be used to identify the topic with highest probability. The model can be employed to identify the similar incidents associated with new incident based on the topic assigned to it. Going forward the model can be used to recommend CAPA for new incident based on the similar incidents.

Appendix

For the Purpose of imparting clarity to the results derived are listed in the following appendix table for the 12 topics along with the corresponding probabilities for the 20 incidences.

TABLE IV: Incident Topic Probability Table

*DNo	Topic	Highest Score	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9	Topic 10	Topic 11	Topic 12
1	Topic 9	0.3475	0.04799	0.04966	0.05345	0.0511	0.048	0.0507	0.15607	0.04843	0.3475	0.0493	0.04982	0.04793
2	Topic 11	0.4579	0.04739	0.04825	0.048	0.0474	0.0491	0.0474	0.06295	0.04895	0.0478	0.04739	0.45794	0.04739
3	Topic 7	0.4241	0.05146	0.04904	0.04917	0.0478	0.0481	0.0479	0.42415	0.04799	0.0487	0.04766	0.05	0.08816
4	Topic 7	0.2671	0.04887	0.04888	0.04889	0.0489	0.0499	0.0497	0.26711	0.04888	0.0516	0.05092	0.04899	0.23736
5	Topic 1	0.4837	0.48371	0.06854	0.05189	0.0451	0.0456	0.0442	0.043	0.04369	0.043	0.04393	0.04376	0.04356
6	Topic 12	0.5329	0.04026	0.04025	0.04068	0.0404	0.0402	0.054	0.04593	0.04044	0.0404	0.04436	0.04016	0.5329
7	Topic 11	0.4738	0.04708	0.04995	0.04883	0.0471	0.0476	0.0471	0.04914	0.04708	0.0482	0.0471	0.47378	0.04708
8	Topic 4	0.4838	0.04043	0.04827	0.07302	0.4838	0.0404	0.0651	0.04676	0.04043	0.0404	0.04043	0.04043	0.04043
9	Topic 11	0.4296	0.04845	0.04877	0.04897	0.0484	0.0484	0.0486	0.04855	0.08401	0.0492	0.04845	0.42964	0.04845
10	Topic 8	0.4179	0.05058	0.05759	0.05093	0.0506	0.052	0.0508	0.05521	0.41785	0.0515	0.05068	0.06168	0.05059
11	Topic 5	0.4341	0.04624	0.04624	0.04814	0.0463	0.4341	0.0462	0.07455	0.04624	0.0547	0.04624	0.04831	0.0627
12	Topic 3	0.4531	0.04526	0.04756	0.4531	0.044	0.0739	0.044	0.04726	0.04757	0.0494	0.04403	0.05992	0.04401
13	Topic 3	0.43	0.0516	0.05213	0.42996	0.0518	0.0528	0.0517	0.05161	0.0516	0.0519	0.0516	0.05175	0.0516
14	Topic 10	0.4604	0.04284	0.09209	0.04335	0.0595	0.043	0.0429	0.04297	0.04391	0.0433	0.4604	0.04292	0.04291
15	Topic 10	0.4865	0.04378	0.04403	0.04432	0.0444	0.044	0.0451	0.04419	0.04401	0.0438	0.48647	0.04378	0.07205
16	Topic 12	0.4468	0.07283	0.08822	0.06157	0.042	0.0415	0.0409	0.04083	0.04292	0.0408	0.04082	0.04082	0.44679
17	Topic 10	0.4517	0.04405	0.04459	0.08696	0.0441	0.0441	0.045	0.04821	0.04522	0.044	0.4517	0.0442	0.05792
18	Topic 4	0.4465	0.04445	0.04775	0.04562	0.4465	0.0444	0.0454	0.04518	0.04453	0.0472	0.08974	0.0547	0.04446
19	Topic 5	0.4091	0.04133	0.0702	0.04428	0.0743	0.4091	0.0599	0.06769	0.04167	0.0414	0.04203	0.04111	0.06693
20	Topic 9	0.3692	0.04739	0.08443	0.04929	0.0479	0.0475	0.0482	0.07518	0.05394	0.3692	0.04826	0.08136	0.04732

*DNo represents Document Number.

ACKNOWLEDGEMENT

The authors thankfully acknowledge the support provided by Dr Varma S Rudraraju, CEO of Aizant Drug Research Solutions Pvt. Ltd, and Sri Chandra Dasika, CTO of Aizant Global Analytics Pvt. Ltd.

REFERENCES

[1] S. Bhutada, "Drug dote and healthcare analysis using topic modeling," Int. J. Innov. Technol. Explor. Eng., vol. 8, no. 8, pp. 1250–1255, 2019.

[2] J. Smith, B. Ghotbi, S. Yi, and M. Parsapoor, "Non-pharmaceutical intervention discovery with topic modeling," arXiv, pp. 2–4, 2020.

[3] W. M. Darling, "A theoretical and practical implementation tutorial on topic modeling and gibbs sampling," Proc. 49th Annu. Meet. Assoc. Comput. Linguist. Hum. Lang. Technol., pp. 642–647, 2011.

[4] M. Kandukuri and HaraGopal. V.V., "Topic Modelling Extraction of 'Mann Ki Baat," Eur. J. Math. Stat., vol. 2, no. 1, pp. 1–12, 2021, doi: 10.24018/ejmath.2021.2.1.11.

[5] N. Aletras and M. Stevenson, "Evaluating topic coherence using distributional semantics," in Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013). Association for Computational Linguistics, 2013, pp. 13-22.

[6] M. Röder, A. Both, and A. Hinneburg, "Exploring the space of topic coherence measures," WSDM 2015 - Proc. 8th ACM Int. Conf. Web Search Data Min., pp. 399–408, 2015, doi: 10.1145/2684822.2685324.

[7] S. Ferilli, F. Esposito and D. Grieco, "Automatic Learning of Lingustic Resources for Stopword Removal and Stemming from Text", in Proceedings of the 10th Italian Research Conference on Digital Libraries (IRCDL 2014), pp. 116-123, 2014, doi: 10.1016/j.procs.2014.10.019.

[8] S. Syed and M. Spruit, "Full-Text or abstract? Examining topic coherence scores using latent dirichlet allocation," Proc. - 2017 Int. Conf. Data Sci. Adv. Anal. DSAA 2017, vol. 2018-January, pp. 165–174, 2017, doi: 10.1109/DSAA.2017.61.