IN-LDA: An Extended Topic Model for Efficient Aspect Mining

Nikhlesh Pathik and Pragya Shukla

Abstract In last decade, LDA is extensively used for unsupervised topic modeling, and various extension of LDA has also been proposed. This paper presents a semi-supervised extension IN-LDA, which uses very few influential words related to the domain for providing supervision in the topic generation process. IN-LDA also improves the performance of the LDA generation process in two ways. First, it deals with multi-aspect terms by passing N-grams vectors, and second simulated annealing-based algorithm is used for tuning hyperparameters of LDA for more coherent output. The experiment is conducted on two popular datasets, movie reviews and 20Newsgroup. IN-LDA is showing improved results when compared with others on coherence value. It also shows a better interpretation of output due to influential words.

Keywords Semi-supervised LDA · Topic modeling · Aspects mining · Hyperparameters tuning · Simulated annealing · Coherence

1 Introduction

Topic modeling first came into focus in 2003 when Beli et al. presented the LDA model for unsupervised clustering from the text in the form of various topics [1]. A large volume of text data can be analyzed with the help of topic models such as LDA. It clusters the documents into various topics. LDA applied to unlabeled data, and it finds clusters of these text documents and grouped them into topics. Different other approaches are applied for aspect extraction. Gou et al. proposed a supervised topic modeling by using the TF-IDF topic frequency method, where the weight of the topic distinguishes the topics. Symmetric and asymmetric Dirichlet prior both have taken as parameters [2]. Shukla et al. proposed various extensions of LDA for efficient aspect extraction from text data [3].

Frequency-based methods fail as different words may be used to represent the same aspect. Rule-based approaches are much dependent on domains and manually

N. Pathik (✉) · P. Shukla
Institute of Engineering and Technology, DAVV, Indore, India

H. Sharma et al. (eds.), *Congress on Intelligent Systems*, Advances in Intelligent Systems and Computing 1334, https://doi.org/10.1007/978-981-33-6981-8_29

designed rules. In this paper, an extended semi-supervised LDA model is presented. The proposed model focused on the issues raised above and tried to make an efficient extension of LDA. The following are the main contributions of the proposed work:-

1. LDA hyperparameter tuning for better coherence value.
2. N-gram vectors for multi-word aspect extraction.
3. Influential words for guiding LDA output and experimentally verify the effectiveness of IN-LDA.

The rest of the paper is organized as follows: Related work is described in Sect. 2. Section 3 discussed the proposed methodology in detail; the experimental setup of the proposed work is explained in detail in Sect. 4. The result and evaluation are discussed in detail in Sect. 5. Section 6 concludes the paper with some future directions.

2 Related Work

Alqaryouti et al. presented a model which aims to help the government to know customers need. They proposed a hybrid approach based on lexicon and semantic rules for aspect extraction and sentiment classification [4]. A supervised extended LDA model was proposed by Zahedi et al. for aspect and sentiment analysis to analyze unlabeled online available opinion datasets [5]. Aziz et al. have used an unsupervised Senti-Wordnet approach for opinion mining. They create a classification model using supervised SVM for defining an opinion [6]. Hughes et al. proposed a training framework for supervised LDA and level prediction. The objective of training generative models is to coherently integrate supervisory signals though a small fraction of training data that are labeled. The model gives high-dimensional logistic regression on text analysis [7].

Fu et al. proposed a topic identification criterion using second-order statistics of the words. This criterion identifies the underlying topics from repulsively violated anchor-word assumptions. They proposed an algorithm for optimization and a primal-dual algorithm [8]. Jabr et al. developed a methodology to represent reviews for each product to measure the reviewer's satisfaction with the extracted aspects. They used machine learning and text mining technique to obtain the product aspects that are important to reviewers from the Amazon dataset [9]. Jin et al. explored an approach to utilize review information for recommendation systems. They named this model as LSTM-Topic matrix factorization (LTMF). LSTM and topic modeling are integrated into the model for review understanding. They proved that LTMF is better for topic clustering than the traditional topic model-based methods [10].

Bagheri proposed a joint generative model sentiment-aspect model (SAM) for aspect mining and sentiment analysis. SAM can detect aspect and their sentiment jointly from the large collection of reviews [11]. Jiang et al. presented a dataset in which each sentence contains multiple aspects with multiple sentiments. They proposed CapsNet and CapsNet-BERT models to learn the complicated relationship between aspects and contexts [12].

Smatana et al. proposed neural networks-based autoencoder known topic AE for topic modeling with input texts. Topic modeling is performed to uncover hidden semantic structures from the input collection of texts. They solved the problem of all three types of topic modeling tasks, e.g., basic topic model, the evolution of topics in time, the hierarchical structure of sub-topics [13]. Bhat et al. proposed two models using deep neural networks. They proposed two variants 2NN Deep-LDA and 3NN Deep-LDA. They have used the Reuters-21578 dataset for experimental evaluation using the support vector machine (SVM) classifier. They explore the modeling of the statistical process of LDA using deep neural network [14].

Luo proposed a text sentiment analysis method based on LDA and CNN to improve the performance of sentiment analysis of public opinion available on the internet. LDA is used for latent semantic representation and CNN as a classifier [15]. Gallagher et al. introduced a correlation explanation approach for topic modeling. This framework separates the most informative topic through anchor words [16]. Jo et al. presented a model with neural network architectures. The model combines two layers of long short-term memory (LSTM) and a latent topic model. The latent topic model trained a classifier for mortality prediction and learned latent topics. They have been designed topic modeling layers for topic interpretability as a single-layer network with constraints inspired by LDA. The LSTM layer captures long-range dependencies in sequential data, trained for mortality prediction [17].

Garcia-Pablosa et al. proposed an unsupervised topic modeling-based model W2VLDA, which performs multilingual and multi-domain aspect-based sentiment analysis [18]. Hai et al. proposed a joint supervised probabilistic model for aspect extraction and sentiment analysis. This extended LDA model represented a review document as opinion pairs and developed an efficient collapsed Gibbs sampling-based inference method for parameter estimation [19]. Lim et al. proposed a model for opinion mining and sentiment analysis based on LDA for twitter opinion data. It influences mentions, emoticons hashtags, and strong sentiment words of tweets [20]. Ramesh et al. explored the use of seeded topic models as Seeded-LDA. It is an extension of topic models that uses a lexical seed set to bias the topics according to relevant domain knowledge. They extract seed words for the COURSE topic from each course's syllabus and capture the sentiment polarity using opinion finder. They explored the massive open online courses (MOOC), a discussion forum platform for student discussions [21]. The majority of research done for improving LDA performance can be summarized on the work done in the following directions:

1. Tuning LDA hyperparameters as a lot of fine-tuning are required [22, 23].
2. Multi-word aspect extraction, because so many aspects, is multi-word.
3. Interpretation of output required human intervention as there is no control over LDA output.

3 Proposed Method

A semi-supervised efficient extension IN-LDA is presented for multi-aspect extraction from text reviews. IN-LDA improves the performance of the LDA topic generation process in two ways. Firstly, it tunes LDA hyperparameter alpha and beta for optimizing LDA performance using simulated annealing-based optimization algorithm SA-LDA. Secondly, for controlling the LDA output, it uses some influential words. N-grams vectors are taken as input to LDA to deal with the multi-word aspect. The detailed description is as follows:

1. The Dirichlet priors α, β significantly affects the LDA performance. SA-LDA algorithm is used for tuning of hyperparameter. SA-LDA works on the principle of simulated annealing and gives the best configuration for the hyperparameter, which increases the performance. Figure 1 represents the flow of the SA-LDA algorithm
2. Aspects are not always unigrams, so for dealing with multi-words aspects, an N-gram approach is used. Bigrams and trigrams are used for the identification of multi-word aspects. The co-occurrence of various words helps in dealing with multi-word and incorrect aspects. It will reduce aspect space, which makes it efficient.
3. We consider only top words (say 100 words) up to three grams. By applying the POS ruleset, we can also find out the main aspects with associated sentiments.
4. There is no control over output in LDA, which makes output interpretation littlie difficult. For these few influential words, say 10–12 words per aspect category is supplied to LDA for some supervision in the topic generation. Aspects are mostly domain-dependent, and sometimes, different words are used for representing the same aspect. Influential words will group these words to identify that aspect. It provides control over topic generation so that co-occurred and relevant words will represent the same topic.
5. The performance of the proposed LDA is compared with LDA and its popular extensions. So, two domains, news and movie, are considered. Top generated words represented the performance improvement based on the coherence value.

Figure 2 represents the flow of the proposed algorithm. Step-by-step description is mentioned below.

3.1 Proposed Algorithms

1. Input: Text review dataset.
2. Preprocessing: Following preprocessing is done on input datasets:
   (a) Tokenization:
      i. Document split into sentences
      ii. Sentences split into words

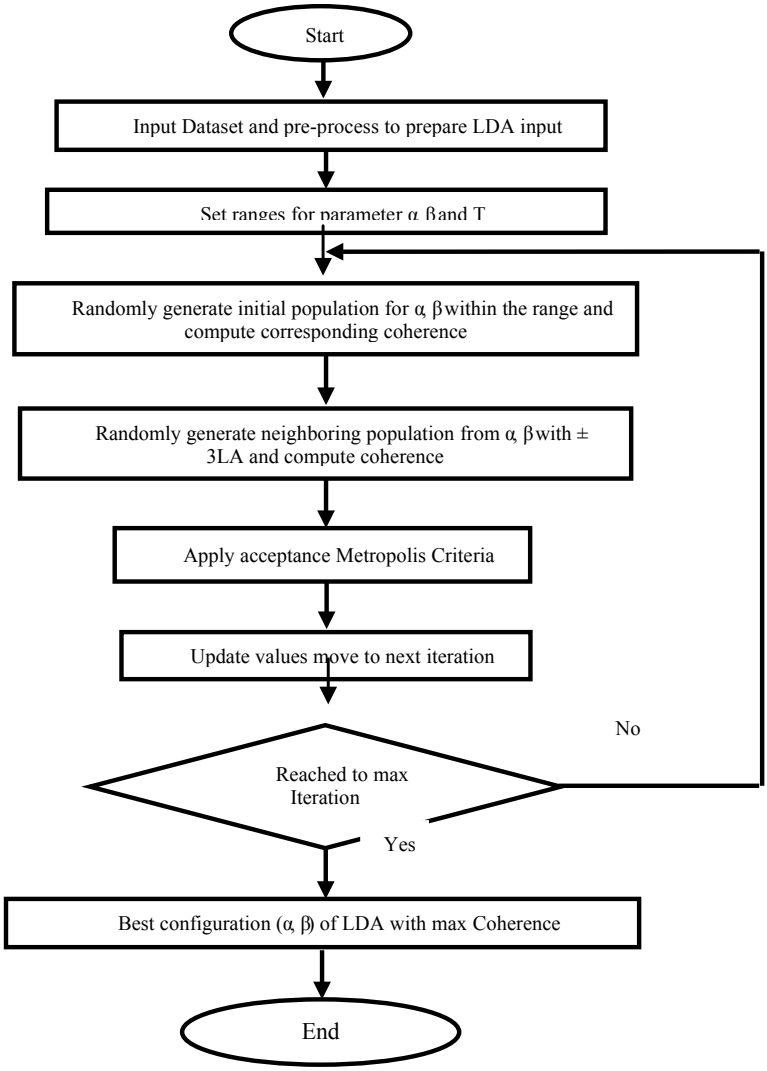

Fig. 1 Flow chart for the SA-LDA algorithm

iii. Words converted into lowercase, and punctuations were removed.

(b) Stop-word elimination: Words like pronouns, preposition conjunction articles, etc. are eliminated. Specific rules are also applied for removing non-useful words. For example, words that have less than three characters are also excluded.

(c) Lemmatization: Similar meaning words are replaced with one word.

(d) Stemming: Words are replaced with their root form.

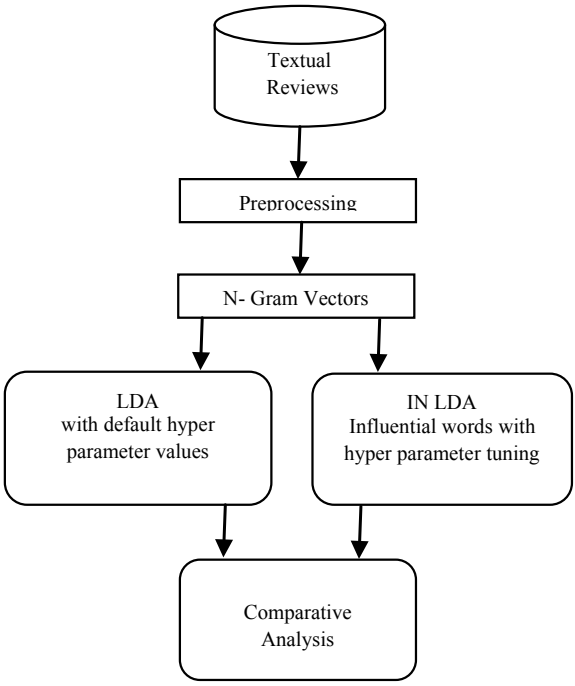

Fig. 2 Flow of proposed work

3. N-gram vectors: Preprocessed data is converted into N-gram vectors word vectors.
4. LDA is applied on step 3 outcome. Topics and their word distribution are obtained.
5. LDA hyperparameter tuning with SA-LDA algorithms.
6. Influential words are identified based on the domain knowledge that the top word list obtained in step 4.
7. IN-LDA applied on step 3 outcome with step (5) parameter values and step (6) influential words.
8. Output: Topics and their word distribution along with the Coherence value.
9. Repeat step1 to 8 with the different datasets for performance evaluation.

4 Experiment Setup

We have considered two popular datasets: a movie review from IMDB and a news feed from 20Newsgroup. The movie review dataset contains 2000 reviews with 1000 positive and 1000 negative. It contains 71,532 sentences with 1,583,688 words. 20 Newsgroup data is a collection of around 20 k news posts on 20 different topics. We supply N-gram input vectors along with influential words for guiding the LDA output.

Table 1 Influential words for the movie review and 20 Newgrouop

Domain	Category	Influential words
Movie	Thriller	Murder, psycho, humor, killer, doubt, crime, crime, dual,
	Comedy	Comedy, laugh, joke, funny, entertain, comic, soft, family, entertain, comic,
	Horror	Horror, scary, devil, death, victim, dark, alien, blood, monster, snake, death,
	Love	Romance, love, life, family, wife, husband, father, mother, son daughter, marriage, Romeo, Juliet
	Music	Music, song, singer, soft, soulful, light, cool, fast, old
	Action	Matrix, Jacki, Chan, war, action, Arnold, Steven
	cartoon	Cartoon, Disney, Mickey, Jerri, king, school, young, effect
News	Sports	Game, team, play, player, hockey, season, score, leagu
	Religion	Hristian, Jesus, exist, Israel, human, moral, ISRA, Bibl
	Violence	Kill, bike, live, leave, weapon, happen, gun, crime, hand
	Graphics	Drive, sale, driver, wire, card, graphic, price, apple, software, monitor
	Technology	File, window, program, server, avail, applic, version, user, entri
	Politics	Armenian, public, govern, Turkish, Columbia, nation, president, group
	Space	space, NASA, drive, scsi, orbit, launch, data, control, earth, moon

Influential words provided supervision to LDA. Table 1 represents the influential words for the movie domain related to various topics.

Around 5–10 words per topic are taken for both movies and news domains. Similarly, for other datasets, influential words will be taken. LDA hyperparameters (α, β) tuning is done using simulated annealing (SA-LDA)-based algorithm. SA-LDA gives the best LDA configuration in terms of hyperparameters values, which provide max coherence value.

Gensim and NLTK LDA implementation is taken as reference and further customized as per the proposed method. The experiment is done on I5 4200 M CPU @ 2.5 GHz processor with 8 GB RAM and Windows 8.1 OS. Anaconda environment is used for evaluation. A simple LDA model is run for finding the topic and word distribution for both unigram and bigram. The output is shown in Table 2.

5 Result and Discussion

IN-LDA produces a more coherent output due to influential words and hyperparameter tuning. Influential words guide LDA to produce more coherent and user interpretable output. For the final LDA output, we have taken ten topics per dataset and fifteen words per topic. Table 3 shows IN-LDA top words for the movie review

Table 2 LDA top words list with unigram for the movie review and 20 Newsgroup

Topic	Top words for movie review	Top words for 20 Newsgroup
Topic 0	film, movie, like, character, time, good, scene, go, play, look, come, story, know, year, thing	line, subject, organ, drive, post, universe, card, write, nntp, host, work, problem, need, article
Topic 1	action, chan, film, role, good, perform, jacki, jack, plot, gibson, team, jone, steven, stone, jam	line, subject, organ, write, nasa, space, article, encrypt, chip, clipper, post, like, know, host, acess
Topic 2	school, life, anim, love, comedy, young, high, play, family, girl, father, elizabeth, beauty, disney, voic	Game, team, line, subject, organ, play, hockey, player, univers, write, post, year, article, host, think
Topic 3	music, story, king, battle, life, ryan, world, prince, spielberg, american, visual, voice, john, song, great	window, line, subject, organ, file, write, post, program, host, nntp, problem, univers, thank, know, graphics
Topic 4	drug, girl, spice, arnold, jerri, park, stiller, young, gorilla, tarantino, tarzan, garofalo, grant, disney, schwarzenegg	Write, israel, article, subject, isra, line, Organ, armenian, people, jew, arab, post, say, kill, think
Topic 5	comedy, funny, laugh, joke, humor, play, comic, hilari, eddi, brook, star, love, gag, amus, american	Chrishtian, write, subject, line, people, organ, think, believe, know, article, jesus, say, univers, like, come
Topic 6	life, love, wife, husband, story, white, crime, town, tell, daughter, death, relationship, henri, perform, Michael	Line, subject, organ, write, article, like, post, bike, host, nntp, look, think, universe, know, good
Topic 7	love, perform, wed, carrey, william, truman, family, life, sandler, wife, play, comedy, marry, role, julia	Write, people, line, organ, subject, article, think, right, post, state, like, govern, nntp, know, host
Topic 8	film, effect, human, star, special, action, planet, world, earth, fight, video, game, science, origin, fiction	Organ, subject, line, article, write, post, know, pitt, bank, food, gordon, science, universe, like, think
Topic 9	alien, horror, movie, vampire, ship, know, crew, killer, kill, origin, scream, summer, scari, sequel, blood	Line, organ, subject, write, year, article, game, post, team, think, nntp, baseball, host, player, universe

and 20Newsgroup dataset. Here, we have only considered single word aspects. For multi-word aspects, we have converted out unigram input into bigram.

Table 3 shows more relevant words on the topics. Influential words supervised the topic generation process and produced a better output. Similarly, for generating multi-word aspects, we have taken N-gram(bigram) input vectors along with the N-gram(bigram) influential words.

Table 4 represents the multi-word (bigrams) aspects generated by IN-LDA when bigram vectors are passed as input. We have only shown top bigram words generated by IN-LDA in the movie and news domains. The output is more interpretable as compare to traditional LDA.

Table 3 IN-LDA unigram output top words for movie and 20 Newsgroup datasets

Topic	Top words for movie review	Top words for 20 Newsgroup
Topic 0	life, love, character, story, live, family, perform, wife, turn, beauty, play, father, relationship, work, begin	line, subject, organ, drive, post, universe, card, write, host, nntp, work, problem, know, article
Topic 1	action, plot, chan, fight, jacki, sequenc, hero, jone, chase, gibson, kill, star, scene, stunt, partner	write, line, subject, organ, article, encrypt, chip, post, clipper, space, like, know, host, nasa, access,
Topic 2	anim, king, voic, disney, story, young, children, family, warrior, prince, mulan, gorilla, kid, snake, adventure	game, line, team, organ, subject, write, year, player, article, play, universe, post, think, host, hockey
Topic 3	girl, music, wed, comedy, school, romantic, julia, love, high, play, singer, song, football, team, band	window, line, subject, organ, file, write, post, host, program, nntp, know, thank, universe, graphic, display
Topic 4	drug, reev, live, play, funny, zero, mike, baldwin, matrix, game, comedy, park, keanu, party, tarantino	write, subject, article, people, organ, line, israel, state, isra, post, think, armenian,say, kill, american
Topic 5	comedy, funny, laugh, joke, humor, play, comic, murphi, get, eddy, carrey, think, moment, brook, jack	write, subject, christian, line, people, organ, think, know, believ, articl, say, post, jesus, like, universe, come
Topic 6	crime, harry, stone, director, murder, francy, cage, investig, mystery, killer, detect, shoot, life, neighbor, blood	line, write, subject, organ, article, like, post, think, host, nntp, know, bike, good, universe
Topic 7	film, perform, character, cast, truman, good, carry, perfect, ryan, direct, john, excel, role, joan, steven	Write, people, line, organ, subject, article, think, right, post, state, like, govern, nntp, know, host
Topic 8	film, movie, like, time, character, good, scene, go, look, know, thing, plot, play, come, think	organ, line, subject, article, write, post, universe, host, know, nntp, reply, pitt, bank, food, think,
Topic 9	alien, horror, vampire, human, ship, planet, effect, crew, earth, origin, special, space, kill, scream, killer	game, line, team, organ, subject, write, year, player, article, universe, post, think, host, team, nntp

Table 4 IN-LDA top multi-word (bigram) generated for movie and 20Newsgroup datasets

Domain	Top multi-words
Movie	high_school, romantic_comedy, true_love main_character, best_friend, funny_movie, real_life, love_story, film_star, action_scene, action_sequence hong_kong, special_effect, action_film, million_dollar, motion picture, star_war, true_stori, support_cast, jam_bond, spice_girl, science_fiction, jurass_park, lake_placid, virtual_realiti, aspect_ratio, jacki_chan, urban_legend, video_game
News	comp_window, date_line, comp_hardwar, video_card, gate_comp, window_misc misc_comp, death_day,,imag_process misc_path, crime_street public_road, window_programm comp_graphic, newsgroup_comp kill_women, window_app, graphic_subject, modem_connect, signal_output

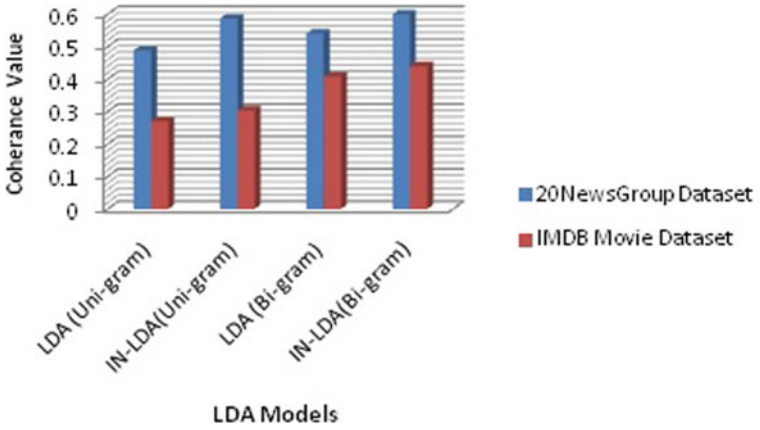

Fig. 3 Coherence plot for IN-LDA with two different datasets

Both LDA outputs are compared based on coherence value, and IN-LDA produced better coherence. With the 20Newsgroup dataset and traditional LDA, value of coherence is 0.488 for unigram, whereas with IN-LDA with the same configuration, coherence is 0.586. When we considered bigram, the value of coherence is 0.54 and 0.60, respectively. Similarly, with movie review, dataset for the unigram value of coherence is 0.27 and 0.304, respectively, for bigram values is 0.407 and 0.44. We can see that influential words guiding LDA for better coherent output. Figure 3 represents the coherence plot for IN-LDA with two above considered datasets.

6 Conclusion

In this paper, semi-supervised IN-LDA is presented for the extraction of multi-word aspects from text reviews. IN-LDA improves the performance of LDA by tuning LDA hyperparameters and providing supervision in terms of influential words. IN- LDA can extract multi-word aspects with better coherence as compare to traditional LDA, which makes output interpretation more clear. The experiment is conducted on two different datasets, which demonstrate the superiority of IN-LDA over LDA. In the future, we will try to combine LDA with neural networks, or a deep learning-based model can be proposed for further performance improvement.

References

1. Blei, D., Carin, L., Dunson, D.: Probabilistic topic models. IEEE Signal Process. Mag. **27**(6), 55–65 (2010)
2. Gou, Z., Huo, Z., Liu, Y., Yang, Y.: A method for constructing supervised topic model based on term frequency-inverse topic frequency. Symmetry **11**(12), 1486 (2019)
3. Pathik, N., Shukla, P.: An ample analysis on extended LDA models for aspect based review analysis. Int. J. Comput. Sci. Appl. **14**(2) (2017)
4. Alqaryouti, O., Siyam, N., Monem, A.A., Shaalan, K.: Aspect-based sentiment analysis using smart government review data. Appl. Comput. Inf. (2019)
5. Zahedi, E., Saraee, M.: SSAM: toward Supervised Sentiment and Aspect Modeling on different levels of labeling. Soft. Comput. **22**(23), 7989–8000 (2018)
6. Aziz, M.N., Firmanto, A., Fajrin, AM., Ginardi, R.H.: Sentiment analysis and topic modelling for identification of government service satisfaction. In: 5th International Conference on Information Technology, Computer, and Electrical Engineering 2018, ICITACEE, pp. 125–130. IEEE (2018)
7. Hughes, M.C., Hope, G., Weiner, L., McCoy, Jr., T.H., Perlis, R.H., Sudderth, E.B., Doshi-Velez, F.: Semi-supervised prediction-constrained topic models. In: AISTATS 2018, pp. 1067–1076 (2018)
8. Fu, X., Huang, K., Sidiropoulos, N.D., Shi, Q., Hong, M.: Anchor-free, correlated topic modeling. IEEE Trans. Pattern Anal. Mach. Intell. **41**(5), 1056–1071 (2018)
9. Jabr, W., Cheng, Y., Zhao, K., Srivastava, S.: What are they saying? A methodology for extracting in-formation from online reviews (2018)
10. Jin, M., Luo, X., Zhu, H., Zhuo, H.H.: Combining deep learning and topic modeling for review understanding in the context-aware recommendation. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies 2018, vol. 1, pp. 1605–1614 (2018)
11. Bagheri, A.: Integrating word status for joint detection of sentiment and aspect in reviews. J. Inf. Sci. **45**(6), 736–755 (2019)
12. Jiang, Q., Chen, L., Xu, R., Ao, X., Yang, M.: A challenge dataset and effective models for aspect-based sentiment analysis. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing 2019, EMNLP-IJCNLP, pp. 6281–6286 (2019)
13. Smatana, M., Butka, P.: TopicAE: a topic modeling autoencoder. Acta Polytech. Hung. **16**(4) (2019)
14. Bhat, M.R., Kundroo, M.A., Tarray, T.A., Agarwal, B.: Deep LDA: a new way to topic model. J. Inf. Optim. Sci., 1–2 (2019)
15. Luo, L.X.: Network text sentiment analysis method combining LDA text representation and GRU-CNN. Pers. Ubiquit. Comput. **23**(3–4), 405–412 (2019)
16. Gallagher, R.J., Reing, K., Kale, D., Ver Steeg, G.: Anchored correlation explanation: topic modeling with minimal domain knowledge. Trans. Assoc. Comput. Ling., 529–42 (2017)
17. Jo, Y., Lee, L., Palaskar, S.: Combining LSTM and latent topic modeling for mortality prediction. arXiv preprint arXiv:1709.02842 (2017)
18. García-Pablos, A., Cuadros, M., Rigau, G.: W2VLDA: almost unsupervised system for aspect-based sentiment analysis. Expert Syst. Appl. **91**, 127–137 (2018)
19. Hai, Z., Cong, G., Chang, K., Cheng, P., Miao, C.: Analyzing sentiments in one go: a supervised joint topic modeling approach. IEEE Trans. Knowl. Data Eng. **29**(6), 1172–1185 (2017)
20. Lim, K.W., Buntine, W.: Twitter opinion topic model: extracting product opinions from tweets by lever-aging hashtags and sentiment lexicon. In: Proceedings of the 23rd ACM International Conference on Information and Knowledge Management 2014, pp. 1319–1328. ACM (2014)
21. Ramesh, A., Goldwasser, D., Huang, B., Daumé, III H., Getoor, L.: Understanding MOOC discussion forums using seeded LDA. In: Proceedings of the 9th Workshop on the Innovative use of NLP for Building Educational Applications, pp. 28–33 (2014)

22. Yarnguy, T., Kanarkard, W.: Tuning Latent Dirichlet Allocation parameters using ant colony optimization. J. Telecommun. Electron. Comput. Eng. (JTEC) **10**(1–9), 21–24 (2018)
23. George, C.P., Doss, H.: Principled selection of hyperparameters in the Latent Dirichlet Allocation model. J. Mach. Learn. Res. **18**(1), 5937–5974 (2017)