A Topical Approach to Capturing Customer Insight Dynamics in Social Media

Miguel Palencia-Olivar[1,2(✉)]

[1] Lizeo IT, 42 quai Rambaud, 69002 Lyon, France
miguel.palencia-olivar@lizeo-group.com
[2] Laboratoire ERIC, Université de Lyon,
5 Avenue Pierre Mendès France, 69500 Bron, France

Abstract. With the emergence of the internet, customers have become far more than mere consumers: they are now opinion makers. As such, they share their experience of goods, services, brands, and retailers. People interested in a certain product often reach for these opinions on all kinds of channels with different structures, from forums to microblogging platforms. On these platforms, topics about almost everything proliferate, and can become viral for a certain time before they begin stagnating, or extinguishing. The amount of data is massive, and the data acquisition processes frequently involve web scraping. Even if basic parsing, cleaning, and standardization exist, the variability of noise create the need for ad-hoc tools. All these elements make it difficult to extract customer insights from the internet. To address these issues, I propose to devise time-dynamic, nonparametric neural-based topic models that take topic, document and word linking into account. I also want to extract opinions accordingly with multilingual contexts, all the while making my tools relevant for pretreatment improvement. Last but not least, I want to devise a proper way of evaluating models so as to assess all their aspects.

Keywords: Topic modeling · Text mining · Social media mining · Business analytics · Natural language processing · Deep learning

1 Introduction

The age of social media has opened new opportunities for businesses. Customers are no longer the final link of a linear value chain; they have also become informants and influencers as they review goods and services, talk about their buying interests and share their opinion about brands, manufacturers and retailers. This flourishing wealth of information located outside traditional channels and frameworks of marketing research also poses many challenges. These challenges, data analysis practitioners must address when trying to test the viability of a business idea, or to capture the full picture and latest trends of consumers' opinion. In particular, social media constitute massive, heterogeneous and noisy document

M. Hagen et al. (Eds.): ECIR 2022, LNCS 13186, pp. 520–527, 2022.
https://doi.org/10.1007/978-3-030-99739-7_64

sources that are often accessed through web scraping when no API is available. Additionally, customer trends tend to evolve through time, thus causing data drifts that create the need for Machine Learning models' updates. Last but not least, documents' structure is oftentimes more complex than in classical applications, as documents can exhibit some linking in the form of a graph (e.g.: Twitter) or in the form of a *nested* hierarchy (e.g.: Reddit). Linking between words and linking between topics are also important for efficient meaning extraction and better interpretation.

To come up with these challenges, I propose to devise time-dynamic, nonparametric neural-based topic models that take topic, document and word linking into account. I also want to address the issue of crosslingual topic modeling, as a single topic can have several versions in different languages. Whatever the modeling decisions, the settings must acknowledge that data is noisy in its founding assumptions, either by filtering it, or by isolating it.

2 Background and Research Questions

My research originates from industrial needs for customer insight extraction[1] from massive streams of texts regarding social media in a broad sense: technical reports, blogs, microblogs (tweets, etc.), and forum posts. Lizeo IT, the company that provides my experimental material, harvests data on a daily basis. This harvesting is mainly performed through web scraping on more than a thousand websites in 6 different languages[2]. The data acquisition pipeline includes parsing and basic data cleaning steps, but noise stills remains, e.g., markup languages, mispells, and documents in a given language that comes from a source supposedly in another language. Due to this noise and to its variability, off-the-shelf tools seldom work. Additionally, no tools are available on specific domains such as the tire industry, which is the main field Lizeo IT evolves in. In-house data dictionaries and ontologies about the tire industries do exist, but they rely on manual, expert knowledge-backed labeling. Lizeo IT's intent for this project is to be able to extract information without any kind of prior background information -objectively observable elements inherent to data set aside- so as to work with data related to other industries. The use case is purely exploratory data analysis (EDA), that is considerably hindered by data opaqueness in absence of background, by data heterogeneity, and by data noisiness.

Consequently, I pursue several goals in terms of both modeling and inference, under precise hypothesis and conditions:

RQ1: Nonparametric topic extraction As the data volumes are huge, one cannot reasonably know what the topics are, nor their number.

RQ2: Integrative extraction I need to extract customers' insights in a way that preserves document, topic and word structures and relationships while taking temporal dependencies, languages and noise into account.

[1] My work solely focuses on the insights per-se, not the emitters, and only includes corpus-related information.

[2] English, french, spanish, italian, german, and dutch.

RQ3: Data cleaning processes improvement I need my approaches to cope with noise directly, either by isolating it, or by filtering it.

RQ4: Scalability The algorithms have to adapt to Big Data-like settings, as the training datasets are massive and as the models need to apply quickly to newly available data.

RQ5: Annotator's bias avoidance Exposed simply, experts, on the one hand, tend to focus on product caracteristics, while on the other hand, customers tend to *focus on their experience of the product*. I want to extract customers' insights to their fullest possible extent.

To solve these problems, I propose fully unsupervised topic modeling approaches. Probabilistic settings in the vein of the Latent Dirichlet Allocation [2] appear as a go-to set of techniques for this task, as they are very flexible statistical models. However, probabilistic topic models are still unsupervised learning techniques that bear an additional -yet desirable- burden: that of *interpretable language modeling*. Analysts have two *simultaneous* expectations: from the purely statistical perspective, they need a model that efficiently extracts latent variables (the topics), that in turn are representative of -or coherent with- the dataset at hand; from the linguistic perspective, they also need *direct* interpretability of the topics. To comply with these expectations:

RQ6: Proper evaluation There is a need for a way of evaluating a topic model by simultaneously taking into account their statistical nature, whatever the modeling or inference settings, and the necessities for interpretability and coherence of the latent variables.

3 Research Methodology

As probabilistic topic models are Bayesian statistics to their very core, it is possible to use all the tools and perspectives the field offers. In particular, it permits to consider distributions as building blocks of a generative process to capture the desired aspects of the dataset at hand. I intend to use *Dirichlet Processes (DPs)* [24] *to automatically determine the number of topics (RQ1)*, and *survival analysis to model their birth, survival and death in time (RQ2)*. The usual Bayesian statistics' workhorse for inference is Markov Chain Monte Carlo (MCMC), but variational inference (VI) [5] has been gaining momentum in the last years due to its ability to scale to massive amounts of data. One of the most recent examples of modern VI is the *Variational Autoencoder framework* [16], *which I intend to use for both scalability (RQ4) and flexibility (RQ1, RQ2)*. The use of black-box variational inference [22] enables fluidifying going through iterative modeling processes such as George Box's modeling loop.

Overall, George Box's modeling loop [3] is useful, as it covers all the expectations one has about modeling tasks. Its iterative structure is simple and software development-friendly[3], and it clearly separates concerns in three distinct steps.

[3] As my work is both statistical and computer-science related, I wanted an exhaustive methodology that could unite both fields as much as possible, with as much emphasis on theoretical concerns as on practical concerns.

The building step corresponds to the actual modeling, when specifications are set so as to capture *the aspects of data that matter to the practitioner*. The freedom the framework allows enables to design parsimonious solutions that give control over the expected results, thus eliminating irrelevant information. This freedom, however, is a double-edged blade that can lead to a validation bias regarding prior, naive assumptions about the dataset at hand. This situation is particularly easily reached in topic modeling due to the very nature of the object of study, and words can be arranged in a way that seem meaningful without them being reflective of the actual corpus. To circumvent the issue, a Bayesian tool comes in handy: *posterior predictive checking (PPC) through realized discrepancies* [14] *. This procedure enables evaluating the whole probabilistic setting of an algorithm while focusing on the important aspects of the model, whatever the underlying probabilistic distributions, thanks to the generative properties of the models (RQ6).* Despite this ability to evaluate a Bayesian setting as a whole, PPCs per-se fail on evaluating the interpretability aspect of topic models. To include semantics -and therefore, interpretability- in the evaluation, *I suggest to regularize the training objective of a VAE with proper regularizers for semantics and coherence* [12] *, so as to twist the inference accordingly with all the objectives, then to perform a PPC (RQ6).* Following this procedure, the PPC should evaluate all the aspects, but the trick is still not enough to integrate all the elements I need to my models.

The hybrid nature of the VAEs, i.e., the fact that these models are neural network-powered probabilistic settings, *allows benefitting from the best properties and advances of both worlds, including embeddings (topics, words, graphs, etc.), transfer learning (RQ1), and recurrent neural networks (temporal aspects, hierarchies, etc.) (RQ1, RQ2)*, for instance. Concerning noise, I believe that it follows Harris' distributional hypothesis, which is the exact one that underlies Mikolov et al.'s works on word embeddings [18]. According to this hypothesis, semantically similar words tend to occur in the same contexts. I think that the statement is also valid for what I call noise, e.g., that code will most likely appear with code if data parsing has failed. *I'm much more inclined to the solution of isolation instead of filtering as it is an additional tool for refining data cleaning processes (RQ3).* As for treating multilingual corpora, embeddings are useful in two ways: the first is that they allow to get distributed representations of words in a classical way[4], as for monolingual contexts; secondly, they allow for the emergence of a pivot language. *A combination between embeddings and adding a categorical distribution in the generative process could help examining crosslingual similarities (RQ2).*

Last but not least, and to circumvent any annotator bias, *I need to restrict knowledge injection to the objectively observable elements: text itself, document structure, time stamps, and the language a document is written in*[5] *(RQ5).*

[4] Except that, words in a given language are much more likely to appear within contexts in the same language.

[5] Language detection is out of the scope of this project, so I either rely on datasets' existing annotations or use off-the-shelf tools.

4 Related Works

Topic extraction is a well established task in the landscape of text mining [6], but the neural flavour of topic models is much more recent, as it originates from the works on black-box variational inference [16,22]. Gaussian VAEs [16] are one of the most famous realization of this line of research. To my knowledge, the *Neural Variational Document Model* (NVDM) [17] is the first VAE-based topic model. Due to its Gaussian prior, however, the NVDM is prone to posterior collapses. To circumvent the issue, Srivastava & Sutton developed the *ProdLDA* [23]. The ProdLDA tries to approximate a Dirichlet prior with a logistic normal distribution to cope with the original reparameterization trick used in the VAE. The ProdLDA makes topic modeling with variational autoencoders stabler, thus highlighting the importance of Dirichlet-like priors in the field. Other Gaussian-based developments include the *TopicRNN* [9], and the Embedded Topic Model (ETM) [10] and its time-dynamic extension [11] (by chronological order). These are particular, in the sense that both train word embeddings directly and conjointly with topic extraction. Prior to these algorithms, an extensive research aimed at linking words in (non-neural) topic models, sometimes while trying to address specific issues such as data mining on short texts [26]. Most of these works treat semantic units as auxiliary information. Other methods switch priors to include some linking between words, tweak word assignment to the topics [8,15,25], or use pre-trained embeddings [1]. Last but not least, and much later, Ning et al. [20] have proposed unsupervised settings that build on stick-breaking VAEs [19], thus automatically finding the number of topics from data.

5 Research Progress and Future Works

My first step for this project has been to adapt the VAE framework to non-parametric settings regarding the number of topics. The DPs have already been applied to topic modeling in non-neural settings. Nalisnick et al. [19] have successfully adapted an approximation of the stick-breaking take on the DPs thanks to a Kumaraswamy variational distribution instead of a Beta for the stick-breaks so as to enforce compliance of the setting with the *original reparameterization trick* [16]. This replacement, however, makes the VAE proner to posterior collapses. To solve the issue, I used Figurnov et al.'s *implicit reparameterization trick* [13] to use a Beta variational distribution, thus making my setting an exact, fully-fledged DP, as shown in [21]. I also included two kinds of embeddings, one for topics, and one for words, to capture similarities in the same embedding space. Finally, I added a Gamma prior on the concentration parameter of my single level DP [4] to learn it from data as it controls the number of topics the model extracts. The whole setting not only efficiently captures similarities and outperforms other state of the art approaches; it also tends to confirm my hypothesis on both crosslingual and noise aspects, as it isolates words by language and noise in separate topics. My approaches were tested on the 20 Newsgroups and on the Humanitarian Assistance and Disaster Relief datasets. On industrial datasets,

I also tried pre-initializing the word embedding matrix with Word2Vec to see if I could get an improvement, without much success regarding getting better results: my settings per-se are at least equivalent to Word2Vec in terms of word embeddings. In the future, I would like to try other kind of prior initializations, with a transformer's embeddings, for instance. Other kinds of future work will mainly consist at improving models' precision, by further improvement of the priors. I'll then move on to adding document linking, with a particular emphasis on nested hierarchies in documents. My last step on the project will consist in adding a categorical distribution in the generative process to capture crosslingual aspects.

I am currently working on extending my models to capture time dynamics, following the approaches of Dieng et al.'s D-ETM [11]. To achieve this, I rely on my models' capacity to use Gamma and Beta distributions to include elements of survival analysis. When its first parameter is set to 1, the Gamma distribution is equivalent to the exponential distribution, which is a common hazard function in survival analysis. Additionally, the Gamma distribution is usable as a conjugate prior for both Gamma distributions with a fixed parameter, and for a DP [4]. As a consequence, the generative process includes a chain of interdependent exponentials (one per time-slice), that will in turn act as conjugate priors for the DPs. I also included the two embedding matrices, but generalized word embeddings to a tensor to get a word distribution per time slice, so as to capture semantic evolution. Besides, I have devised a regularization term to encourage the training procedure to compute neatly separated topics, as per what I learnt on how humans interpret topic models from Chang et al.'s work [7]. I intend to use PPC to assess the whole setting.

6 Specific Research Issues for Discussion at the Doctoral Consortium

The specific research issues I would like to discuss are the following:

- **Issue 1:** The way probabilistic topic models treat corpora end up with a single topic-wise word distribution for all documents in the corpus. However, documents put varying emphasis on words. I would like to "personalize" word distributions with respect to the documents I treat by adding document-wise information, without losing the ability to generalize and predict about future documents.
- **Issue 2:** Some corpora's structures are complex, and modeling through means of a graph or of a hierarchy can help, but some sources can be both at the same time. It is particularly true for Twitter, where users can link to another tweet through retweets and start discussion threads.

References

1. Batmanghelich, K., et al.: Nonparametric spherical topic modeling with word embeddings. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pp. 537–542. Association for Computational Linguistics, Berlin, August 2016. https://doi.org/10.18653/v1/P16-2087
2. Blei, D., et al.: Latent Dirichlet allocation. J. Mach. Learn. Res. **3**, 993–1022 (2003)
3. Blei, D.M.: Build, compute, critique, repeat: data analysis with latent variable models. Ann. Rev. Stat. Appl. **1**(1), 203–232 (2014)
4. Blei, D.M., Jordan, M.I.: Variational inference for Dirichlet process mixtures. Bayesian Analysis **1**(1), March 2006. https://doi.org/10.1214/06-BA104
5. Blei, D.M., et al.: Variational inference: a review for statisticians. J. Am. Stat. Assoc. **112**(518), 859–877 (2017)
6. Boyd-Graber, J., et al.: Applications of topic models. Found. Trends Inf. Retrieval **11**, 143–296 (2017)
7. Chang, J., et al.: Reading tea leaves: how humans interpret topic models. In: Bengio, Y., Schuurmans, D., Lafferty, J., Williams, C., Culotta, A. (eds.) Advances in Neural Information Processing Systems, vol. 22. Curran Associates, Inc. (2009)
8. Das, R., et al.: Gaussian LDA for topic models with word embeddings. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pp. 795–804. Association for Computational Linguistics, Beijing, July 2015
9. Dieng, A.B., et al.: TopicRNN: a recurrent neural network with long-range semantic dependency. arXiv:1611.01702 [cs, stat], February 2017
10. Dieng, A.B., et al.: Topic modeling in embedding spaces. Trans. Assoc. Comput. Linguistics **8**, 439–453 (2020)
11. Dieng, A.B., Ruiz, F.J.R., Blei, D.M.: The dynamic embedded topic model. CoRR abs/1907.05545 (2019)
12. Ding, R., Nallapati, R., Xiang, B.: Coherence-aware neural topic modeling. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. pp. 830–836. Association for Computational Linguistics, Brussels, Belgium, October–November 2018
13. Figurnov, M., others: Implicit reparameterization gradients. In: Proceedings of the 32nd International Conference on Neural Information Processing Systems, NIPS 2018, pp. 439–450. Curran Associates Inc., Red Hook (2018)
14. Gelman, A., Meng, X.L., Stern, H.: Posterior predictive assessment of model fitness via realized discrepancies, p. 76
15. Hu, Y., et al.: Interactive topic modeling. Mach. Learn. **95**(3), 423–469 (2014)
16. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. CoRR (2014)
17. Miao, Y., et al.: Neural variational inference for text processing. In: Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48, ICML 2016, pp. 1727–1736. JMLR.org (2016)
18. Mikolov, T., et al.: Efficient estimation of word representations in vector space. In: Proceedings of Workshop at ICLR 2013, January 2013
19. Nalisnick, E.T., Smyth, P.: Stick-breaking variational autoencoders. In: 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, 24–26 April, 2017, Conference Track Proceedings. OpenReview.net (2017)

20. Ning, X., et al.: Nonparametric topic modeling with neural inference. Neurocomputing **399**, 296–306 (2020)
21. Palencia-Olivar, M., Bonnevay, S., Aussem, A., Canitia, B.: Neural embedded Dirichlet processes for topic modeling. In: Torra, V., Narukawa, Y. (eds.) MDAI 2021. LNCS (LNAI), vol. 12898, pp. 299–310. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-85529-1_24
22. Rezende, D.J., Mohamed, S., Wierstra, D.: Stochastic backpropagation and approximate inference in deep generative models. In: Xing, E.P., Jebara, T. (eds.) Proceedings of the 31st International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 32, pp. 1278–1286. PMLR, Bejing, 22–24 June 2014
23. Srivastava, A., Sutton, C.: Autoencoding Variational Inference For Topic Models, p. 12 (2017)
24. Teh, Y.W., et al.: Hierarchical Dirichlet processes. J. Am. Stat. Assoc. **101**(476), 1566–1581 (2006)
25. Xun, G., et al.: A correlated topic model using word embeddings. In: Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, pp. 4207–4213 (2017)
26. Yan, X., et al.: A biterm topic model for short texts. In: WWW 2013 - Proceedings of the 22nd International Conference on World Wide Web, pp. 1445–1456 (2013)