Aspect-term Extraction from Albanian Reviews with Topic Modeling Techniques

Majlinda Axhiu and Azir Aliu

Southeast European University, Tetovo, North Macedonia
majlinda.axhiu@gmail.com
azir.aliu@seeu.edu.mk

Abstract. Bearing in mind the exponential increase of online data generated by the social networks' users in every language, the urge need of sentiment analysis is also increasing. However, we have reached to a point that even the overall sentiment of an opinion is not enough that is why the necessity of Aspect-based Sentiment Analysis (ABSA) is very high. Considering our aim, to work on the first phase of the ABSA task, namely to extract the aspect terms from the reviews in Albanian language, and considering the lack of research on this field for this language and the lack of resources, we have chosen the unsupervised approach beside the supervised one. In this technique two of the mostly used models that are considered to be the state of art for topic modeling are Latent Dirichlet Allocation (LDA) and Non-negative Matrix Factorization (NMF). We have done a comparative analysis for these two models by using a dataset that we have created from Facebook reviews, in the domain of restaurants. We have successfully extracted the aspects with both models. As a sample of the results we have listed the top 10 words that were extracted by both models and which were classified in three different topics.

Taking into account the results from the evaluation measures (Precision, Recall and F1-score) it resulted that both models worked well for extracting the aspects, having NMF with a higher accuracy than LDA. NMF was also more accurate in the classification of the aspects into different topics.

Keywords: Non-negative matrix factorization · Latent Dirichlet allocation · Aspect extraction · Aspect-based sentiment analysis

1 Introduction

Nowadays people like and insist to make their opinions public and available in the Internet. Their opinions refer to different topics, products and services, social events, individuals, organizations, etc. Considering the increase of e-commerce usage, the number of online reviews is also exponentially increasing. This has led to a situation that almost every potential customer check for other's opinion before buying a product or using a service. On the other perspective, the organizations also want to observe how their products or services are accepted and commented, so that they can modify and adjust their services/products according to the customers' needs.

However, reading all the reviews and trying to make a good decision is very time consuming and almost impossible. Taking into account the huge number of different opinions and sometimes contradictory ones, can result to confusion. That is why the automated sentiment analysis has reached a wide range of application. Sentiment analysis is a set of methods, usually implemented in computer software, that perceive, measure, report and utilize attitudes, opinions, and emotions, which generally are called sentiments and can be found in online, social, and enterprise information sources [2].

When focusing on just the overall sentiment was not sufficient for making a good decision than the research has been broaden and the aspect-based sentiment analysis has been introduced. Aspect-Based Sentiment Analysis (ABSA) refers to systems that determine the opinions or sentiments expressed on different features or aspects of the products or services under evaluation. Namely it aims to extract the aspects of an opinion and to predict the rating of each detected aspect. Aspect terms are attributes or features of a specific topic and ratings are the evaluated interpretation of user satisfaction in terms of numerical values [1].

An ABSA task typically involves two sub-tasks, including identifying relevant entities and aspects, and determining the corresponding sentiment/polarity. In this paper our focus is in the first sub-task: aspect extraction from customer reviews.

There exist three major approaches that are broadly used for aspect detection: supervised, semi-supervised and unsupervised. For the first approach it is required a pre-labelled training dataset that is very expensive to build and requires much human labor. Since finding labeled data for every topic and for each language is very difficult, not to say impossible; always models that work with unstructured and unlabeled data are mostly preferred. Namely, the unsupervised approach is used more when we want to build models that are adoptable for different topics and different languages. And the semi-supervised approach uses a relevant amount of seed words for topic extraction and for sentiment classification.

Since we are analyzing texts in Albanian language and because there are no publicly available labeled data in Albanian language, we chose the unsupervised approach for detecting the aspects. We focus on the topic modeling technique, which is an unsupervised learning approach for clustering documents and discovering topics. The two mostly used topic modeling techniques are: Latent Dirichlet Allocation (LDA) and Non-Negative Matrix Factorization (NMF). LDA uses a probabilistic approach whereas NMF is a linear-algebraic model. In this paper we use both of the mentioned models and we compare the accuracy of the results in the term of top-selected aspects and evaluate their F1 scores.

Both of the models (LDA and NMF) are described in the section 3 while the details for the dataset and the comparative analysis are described in sections 4 and 5 respectively.

2 Related Work

Currently available approaches for the tasks of ABSA can be divided into supervised learning approach [3], semi-supervised [4] or weakly-supervised [5][6] learning approaches, and unsupervised learning approach [7]. For the first technique the most commonly used approaches are: Support Vector Machine, Naïve Bayes classifier, Maximum Entropy classifier and lately Neural Networks. The large number of labeled corpora that these techniques need, in order to perform well, it is an issue for the languages that are low in resource and sometimes even for the other languages but for specific domains. That is why in these cases the weakly-supervised or unsupervised learning approaches are used.

In unsupervised approach no training data is needed. For the extraction of aspects and sentiments in this technique, the topic models are used. The challenge in this stage is to extract both explicit and implicit aspects that may be found in the reviews. Since in the reviews usually we find more informal text, the occurrence of implicit aspects is increasing considerably. They cannot be identified by simple syntactic analysis [8] that is way it is done a lot of comparative analysis among the topic modelling algorithms. The mostly used techniques that are considered state of the art for topic modelling are Latent Dirichlet Allocation and Non-negative Matrix Factorization.

They could both automatically extract the topics from reviews, but they model the texts in different ways; LDA uses the probabilistic generative process, while NMF uses geometrically linear combinations. Therefore, it is challenging and in the interest of researchers to do a comparative analysis of both models. [8]

Current topic modelling approaches are computationally efficient and also seem to capture the correlations between words and topics, but they have two main limitations. The first limitation is that they assume that words are generated independently of each other. This is known as the bag-of-words assumption. In other words, topic models only extract unigrams for topics in a corpus. The second limit for current topic modelling approaches is the assumption that the order of words can be ignored. [9].

Considering the Albanian language, in general there is a lack of research in the field of sentiment analysis. Until now [10] has worked on supervised learning approach, but for unsupervised learning approach for ABSA, to the best of our knowledge there is no research. Thus, our focus on this paper is in the extraction of aspect terms with unsupervised techniques.

3 Description of Topic Modeling Algorithms

3.1 Latent Dirichlet Allocation (LDA)

Latent Dirichlet Allocation (LDA) is a probabilistic model and is considered one of the mostly used models for topic modeling. It uses two probability values, such as P(word/topics) and P (topics/documents).

Initially all words are assigned randomly to several topics. Then with an iterative procedure the probabilities are calculated multiple times, until the convergence of the algorithm [11]. The larger the number of iterations, the better the classification of the topics is done, because successively we will have more updated information for the documents.

So, the main idea is that documents are represented as mixtures of latent topics that contain words, which in aspect-based sentiment analysis are considered as aspects, with specific and corresponding probability of their occurrence.

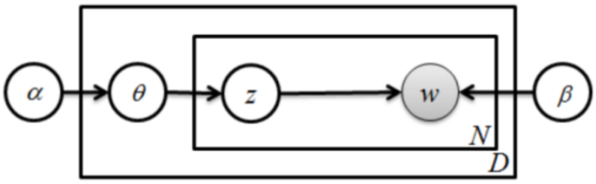

Fig. 1. The graphical model of LDA

In Fig. 1 [1] it is shown the graphical model of LDA. The variables are represented by the nodes, while the dependencies are noted by the edges. Among the nodes, as it can be noticed there are shaded nodes, which represent the observed variables, and not shaded which represent the latent variables. Also, there are two plates which refer to the iterations. The inner plate includes the words, while the outer one includes the reviews. D indicates the number of reviews and N indicates the number of the words in each review.

When we have a corpus of documents with size D, each word w is linked with a latent topic z. From a Dirichlet distribution Dir(α), it is drawn a random sample representing the topic distribution (noted as θ) of a specific document. From a multinomial θ foreach word w, a specific topic z is selected. After that from another Dirichlet distribution of BETA, a random sample which represents the word distribution of topic z is selected. And then from that distribution the word w is chosen. [12]

In a basic LDA model the hyper-parameters (noted as α and β) are symmetric because no previous information about the topic and word distributions is assumed. [5]

With the usage of LDA, as a topic modelling approach, we can extract human-interpretable topics from a dataset, where each topic is characterized by the words, they are most strongly associated with. So, the dataset has the role of the training data for the Dirichlet distribution of document-topic distributions. Even if we haven’t seen a document, we can easily sample from the Dirichlet distribution and move to the next steps from there [4].

Since for the anonymous topics generated from LDA it is needed an additional step to map them to a meaningful topic’s category, some use manual reviews by a professional and the others do a mapping calculation to existing resources [5]. [5] uses the approach of seed words in combination with word embeddings for semantic word similarity.

So, for the further steps of aspect-based sentiment analysis LDA is combined with other algorithms, tasks and models in order to extract the aspects and successfully separate them from the sentiments and proceed with their classification.

3.2 Non-negative Matrix Factorization

Non-Negative Matrix factorization (NMF) is a linear algebraic optimization algorithm. It is considered to be one of the top mostly used models for topic modelling, because it has a huge application in natural language processing. Namely, it can extract meaningful topics in documents without any previous knowledge of the meaning of the data.

The mathematical objective of NMF is to decompose a single input matrix, noted as A, into two non-negative matrices W and H, such that their product is a close estimate to the input matrix. For topic modeling, the input matrix of choice is the document-word matrix [13]. The matrix W represents the topics, namely the clusters that are extracted or discovered from the documents, and the matrix H includes the coefficient weights for the topics in each document [11]. W and H are calculated by optimizing over an objective function and both of the matrices are updated by iteration until convergence.

$$\frac{1}{2}\left|\left|A - WH\right|\right|_F^2 = \sum_{i=1}^{n} \sum_{j=1}^{m} \left(A_{ij} - (WH)_{ij}\right)^2 \qquad (1)$$

In the above written objective function, the error of reconstruction between A and the product of its factors W and H is measured, based on Euclidean distance. By using the objective function, the update rules for matrices W and H can be derived as follows:

$$W_{ic} \leftarrow W_{ic} \frac{(AH)_{ic}}{WHH_{ic}} \qquad (2)$$

$$H_{cj} \leftarrow H_{cj} \frac{(WA)_{cj}}{(WWH)_{cj}} \qquad (3)$$

The recalculation of the reconstruction error is done by using the newly updated values of W and H. This process is repeated until convergence.

The loss of some mathematical precision due to the non-negativity restriction is compensated by a meaningful and coherent representation. [14]

Considering both the explicit and implicit aspects, NMF tends to be mostly preferred for the implicit ones, because it has better results than other algorithms for topic modelling, especially from LDA. [15]

4 Dataset and Preprocessing

The dataset that we used for the experimental part was created from gathering online reviews in Albanian language, from social media- mainly from Facebook. The reviews correspond to the domain of restaurants and in total there are 1015 reviews. The data scraping is done by the tool Facepager [16], with which there were extracted all the comments from restaurants' fun pages, together with their IDs and the corresponding date and time. The comments have a length of 5 to 180 characters.

Taking into consideration that the data was not clean we had to do a considerable amount of data cleaning. In the first stage we removed the people's tags, which were present in a huge amount. After that we also had to remove the symbols and emoticons. Some of the symbols were in the category that we had to reformulate them, e.g. the currencies. After we gained the cleaned data, the following process was the tokenization. In this case we have done a unigram tokenization. During this process it was also important to have ignored or previously cleaned the allowed symbols or punctuation marks that were used with a space. In the following Fig. 2 it is shown the complete flow of data preprocessing.

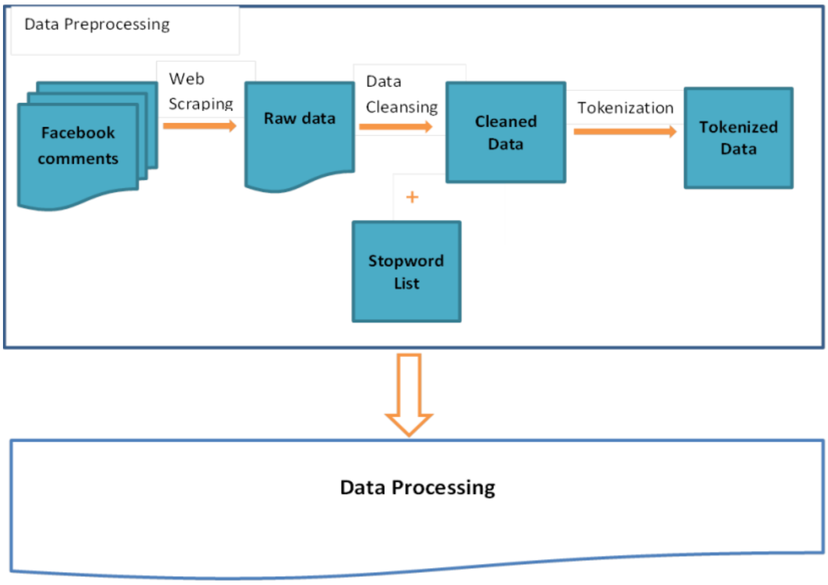

Fig. 2. Data Preprocessing Flow

In the phase of preprocessing we also have built a list of stop-words, in which there are included the conjunctions, determiners, prepositions and other very common words used in Albanian language. We use this list (which consists more than 150 words), in order to ignore those words in the following ABSA tasks, namely in both aspect extraction and sentiment classification.

Both, the dataset with the restaurant reviews and the list of the stop words, will be publicly available soon.

5 Experimental Results

In both models (LDA and NMF) as inputs are provided the reviews and the list of stop-words, and in both of them the number of topics is specified.

The models associate the words to each topic by an iterative procedure, and in this case, we have done 50 iterations. In the following tables there is shown a sample of the results of aspect extraction. In both models there were extracted more than 250 words which were actually considered as aspect terms.

Table 1. Aspects extracted with LDA model with 50 iterations

Sr.No.	Topic 1	Topic 2	Topic 3
1	Ushqim (food)	Shkëlqyeshëm (excellent)	Kamarier (waiter)
2	Kajmak (cream)	Shërbim (service)	Shumë (many)
3	Respekt (respect)	Restoran (restaurant)	Pastër (clean)
4	Fantastik (fantastic)	Ambient (ambiance)	Përvojë (experience)
5	Yje (stars)	Qytet (city)	Mënyrë (method)
6	Keq (bad)	Mrekullueshme (wonderful)	Aromë (odor)
7	Butë (soft)	Patate (potatoes)	Këndshëm (nice)
8	Restoran (restaurant)	Gjatë (long)	Çmim (price)
9	Kamarier (waiter)	Tmerrshëm (terrible)	Mire (good)

10	Speca (peppers)	Kohë (time)	Sukses (success)

In Table 1 there are displayed the top 10 aspects for three different topics, generated by LDA model, while in Table 2 the generated aspects are from NMF model.

Since in both models we haven't used additional algorithms for separation of aspect and opinion (sentiment) it is obvious that there is a considerable amount of words that are wrongly categorized as aspects (noted with red color). However, despite this fact the precision in the first top 10 is around 60-70% in both models. We had a proportionally similar situation also when we extracted 20 or more aspects.

Table 2. Aspects extracted with NMF model with 50 iterations

Sr.No.	**Topic 1**	**Topic 2**	**Topic 3**
1	Ambient (ambiance)	Viçi (beef)	Darkë (dinner)
2	Atmosfera (atmosphere)	Ushqim (food)	Efikas (efficient)
3	Zhurmë (noice)	Mish (meet)	Staf (staff)
4	Bukur (beautiful)	Pica (pizza)	Jashtëzakonshme (extraordinary)
5	Dekor (decoration)	Pikërisht (exactly)	Menaxheri (manager)
6	Çmim (price)	Pulë (chicken)	Menu (menu)
7	Dysheme (floor)	Qepë (onion)	Shërbim (service)
8	Elegant (elegant)	Shkëlqyeshëm (Excellent)	Njerëz (people)
9	Fresket (fresh)	Thatë (dry)	Pasjellshëm (rude)
10	Oborri (garden)	Tryezë (table)	Pronar (owner)

As it can be noticed the number of wrongly assigned words is equal or less in the NMF model. But this is not a sufficient indicator to compare and choose which one is more appropriate and with higher accuracy. That is why we have evaluated the results (in the next section) by the Precision, Recall and F1 measures. However, it is easily noticed that the classification of aspects into topics is better in NMF model, since in LDA we have group of words that should correspond to one topic and here are classified in more than one (e.g. restoran, kamarier, etc.)

In order to evaluate the accuracy of the models and to compare them, we have additionally manually annotated the dataset. The annotation process has been double checked by two different annotators. The number of different aspects in all reviews is 346.

In order to calculate the precision, recall and F1 score of both models we use the following formulas:

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \tag{4}$$

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \tag{5}$$

and

$$F1\ score = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{6}$$

True positives are words classified as aspect terms by the models that actually are aspects (meaning they are correct), false positives are cases the models incorrectly labels as aspects that are actually not (they may be sentiments or other words), and false negatives are words that the models do not identifies as aspects but actually they are.

The results of the measures for both models are shown in Table 3.

Table 3. Results of the evaluation of LDA and NMF

	Precision %	**Recall %**	**F1 score %**
LDA	47.46%	40.46%	43.68%
NMF	50.79%	43.73%	46.99%

Considering all three measures it can be noticed that mainly both of the models provided satisfying results for extracting the aspects, however Non-negative Matrix Factorization model has a better accuracy in comparison to Latent Dirichlet Allocation.

6 Conclusion and Future Work

In this paper we have done a comparative analysis of two mostly used algorithms in the scope of topic modelling, LDA and NMF. They can both automatically extract the topics from reviews, but as we previously saw they model the texts in different ways; LDA uses the probabilistic generative process, while NMF uses geometrically linear combinations. We have tested both of them in the same dataset, which was consisted on 1015 reviews in the domain of restaurants.

Our goal was to analyze which one fits best for texts in Albanian language. And from the tests that we have made the results showed us that both of the models work well for the process of aspect extraction, but NMF has a slightly higher accuracy than LDA. What is important to be noticed is that the classification of the terms into topics is also better with the application of NMF.

We should not ignore the fact the reviews are in informal language and the usage of abbreviations, dialects and slang language can considerably affect the results. Considering this, it is one more time proved that NMF is more suitable when we have to deal with data that have ambiguities and semantic gap.

If we consider the next phase of Aspect based Sentiment Analysis this work can be extended by combining these models with other algorithms and techniques in order to achieve the most accurate separation of aspects and sentiments.

References

1. Moghaddam, S., Ester, M.: On the design of IDA models for aspect-based opinion mining. Proceedings of the 21st ACM international conference on Information and knowledge management - CIKM '12. doi: 10.1145/2396761.2396863 (2012)
2. Liu, B.: Sentiment Analysis and Opinion Mining. Synthesis Lectures on Human Language Technologies 5:1-167. doi: 10.2200/s00416ed1v01y201204hlt016 (2012)
3. Toh, Z., Su, J.: NLANGP at SemEval-2016 Task 5: Improving Aspect Based Sentiment Analysis using Neural Network Features. Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016). doi: 10.18653/v1/s16-1045 (2016)
4. Xiang, B., Zhou, L.: Improving Twitter Sentiment Analysis with Topic-Based Mixture Modeling and Semi-Supervised Training. Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). doi: 10.3115/v1/p14-2071 (2014)
5. García-Pablos, A., Cuadros, M., Rigau, G.: W2VLDA: Almost unsupervised system for Aspect Based Sentiment Analysis. Expert Systems with Applications 91:127-137. doi: 10.1016/j.eswa.2017.08.049 (2018)
6. Purpura, A., Masiero, C., Susto, G.: WS4ABSA: An NMF-Based Weakly-Supervised Approach for Aspect-Based Sentiment Analysis with Application to Online Reviews. Discovery Science 386-401. doi: 10.1007/978-3-030-01771-2_25 (2018)
7. Brody, S., Elhadad, N.: An Unsupervised Aspect-Sentiment Model for Online Reviews. Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics 804–812 (2010)

8. Chen, Y., Zhang, H., Liu, R., et al.: Experimental explorations on short text topic mining between LDA and NMF based Schemes. Knowledge-Based Systems 163:1-13. doi: 10.1016/j.knosys.2018.08.011 (2019)
9. Bagheri, A., Saraee, M., de Jong, F.: ADM-LDA: An aspect detection model based on topic modelling using the structure of review sentences. Journal of Information Science 40:621-636. doi: 10.1177/0165551514538744 (2014)
10. Biba, M., Mane, M.: Sentiment Analysis through Machine Learning: An Experimental Evaluation for Albanian. Advances in Intelligent Systems and Computing 195-203. doi: 10.1007/978-3-319-01778-5_20 (2014)
11. Chawla, R.: Topic Modeling with LDA and NMF on the ABC News Headlines dataset. In: Medium.https://medium.com/ml2vec/topic-modeling-is-an-unsupervised-learning-approach-to-clustering-documents-to-discover-topics-fdfbf30e27df (2017)
12. Xu, J.: Topic Modeling with LSA, PLSA, LDA & lda2Vec - KDnuggets. In: KDnuggets. https://www.kdnuggets.com/2018/08/topic-modeling-lsa-plsa-lda-lda2vec.html (2020)
13. Suri, P., Roy, N.: Comparison between LDA & NMF for event-detection from large text stream data. 2017 3rd International Conference on Computational Intelligence & Communication Technology (CICT). doi: 10.1109/ciact.2017.7977281 (2017)
14. Ho, N.: NonNegative Matrix Factorization Algorithms and Applications. Ph.D, UNIVERSITÉ CATHOLIQUE DE LOUVAIN (2008)
15. Xu, Q., Zhu, L., Dai, T., et al.: Non-negative matrix factorization for implicit aspect identification. Journal of Ambient Intelligence and Humanized Computing. doi: 10.1007/s12652-019-01328-9 (2019)
16. Jünger, J., Keyling, T.: Facepager. An application for automated data retrieval on the web. Source code and releases available at https://github.com/strohne/Facepager/ (2019)