Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 00 (2022) 000–000

Procedia Computer Science

www.elsevier.com/locate/procedia

11th International Young Scientist Conference on Computational Science

# A study of the influence of news reports and other contextual open-source information on the consumer behavior of bank card users

Grigoryev Mikhail[a,*]

[a]ITMO University, Kronverksky Pr. 49, bldg. A, St. Petersburg, 197101, Russia

Abstract

We live in a world filled with media. Thus, a hypothesis can be formulated that background context such as news and open-source statistics can influence the way we consume goods in various categories. This research is an endeavour into the impact of context information on consumer behavior. The study of such influence is performed as prediction of consumption time series with context as exogenous variables.



Keywords: Machine Learning ; Natural Language Processing ; Topic Analysis ; Autoregressive Models ; ARIMAX

## 1. Introduction

## 2. Related work

### 2.1. Topic analysis

The main goal of this part of literature review is the choice of topic analysis methods to be used in this research work for processing news headlines. Thus, the most acknowledged methods ranging from classic LDA to modern BERTopic are looked upon in this section.

Topic analysis is one of the tasks of natural language processing (NLP). It consists of embedding human-legible texts into machine-readable embeddigs, dimension reduction and further clustering of the embeddings. Such clusters can be interpreted as topics found in the texts, hence the name ”topic analysis”. One of the classical methodologies of topic analysis is the ”bag-of-words”, an approach for viewing documents

* Corresponding author. Tel.: +7-921-592-4920.
E-mail address: mikegrig@inbox.ru

as unstructured vocabularies of words with different frequencies. It was the foundation for a plethora of commonly used topic analysis tools such as LDA, NMF, Top2Vec and CTM.

Authors of the publication [4] extracted topics from scientific literature via custom two-stage pipelines. The first step in every experiment was to embed the text into machine-interpretable numeric vectors. Then in the second step different algorithms were used to cluster the vectors, validate the extracted clusters and output an array of topics.

For embedding scientific texts several approaches including word2vec, POS2vec, word-position2vec and LDA2vec were used individually as well as combined into the improved word embedding based scheme. The latter model takes scientific abstract sentences as input, which is processed by word2vec, a simple embedder that transforms all words within the sentence into vectors without taking the word position into account. Then the authors computed averages of word vectors within each sentence so that the resulting vectors could be used as sentence representations. Before averaging, stop-words have been removed due to this tactic yielding better results. After that, POS2vec (part-of-speech2vec) was used on each word in the sentence to generate POS tags. Then, word-position2vec was employed to generate position vectors for the words and the whole sentence. The latter was concatenated to the vectors obtained earlier. Lastly, the sentences were further processed by LDA2vec. Its output vector was concatenated to yield the final ensemble embedding.

In the publication K-means, K-modes, K-means++, SOM and DIANA were chosen as the clustering algorithms. Classical K-means is a simple partition-based clustering method, while K-modes is its extension better suited for categorical data. As K-means' performance highly depends on the initial random seed for the coordinates of cluster centers, K-means++ uses a heuristic function to determine the optimal initial cluster centers. Self-organizing maps (SOM) method is based on clustering and dimension reduction. Divisive analysis clustering (DIANA) is a divisive hierarchical clustering method. Finally, the authors proposed iterative voting consensus (IVC) – a consensus function which seeks cohesion between different clustering methods on the same data and thus lets those algorithms compensate for each others' weaknesses.

The authors have compared 43 different configurations of topic analysis pipeline on the same dataset of scientific abstracts. The models were compared by values of the following metrics: Jaccard Coefficient (JC), FM and F1.

$$JC = \frac{a}{a+b+c} \tag{1}$$

$$FM = \left(\frac{a}{a+b} \cdot \frac{a}{a+c}\right)^{0.5} \tag{2}$$

$$F1 = \frac{2a^2}{2a^2 + ac + ab} \tag{3}$$

Here $a$ denotes the number of pairs of instances which are grouped in the same cluster and fall in the same category in the initial dataset. The value $b$ stands for pairs from the same category which were not grouped in the same cluster. Lastly, $c$ denotes pairs from different categories which were grouped together. The higher the metrics described above, the better results models produce.

According to the results presented in the article and in figure 1 ([4]), among single method clustering strategies, LDA (latent Dirichlet allocation) outperformed other algorithms, with DIANA taking the second place. Results of clustering seemed to have improved significantly when integrating LDA2vec embedding scheme with clustering algorithms.

Even better results were shown by the proposed clustering ensemble framework which utilizes IVC for combining the mentioned methods. Thus, such proposed ensemble approach can lead to better topic analysis models.

As opposed to the example presented above, topic analysis can solve not only the problem of extracting broad semantic themes from texts but also some more nuanced information. Authors of the article [3] propose their two-step approach for aspect-based sentiment analysis. Their procedure lets evaluate reviews in terms of how positive or negative they are and also define aspects of the review object which are evaluated. In the publication, restaurant and laptop reviews were used as input data. The proposed model outputs a tuple in

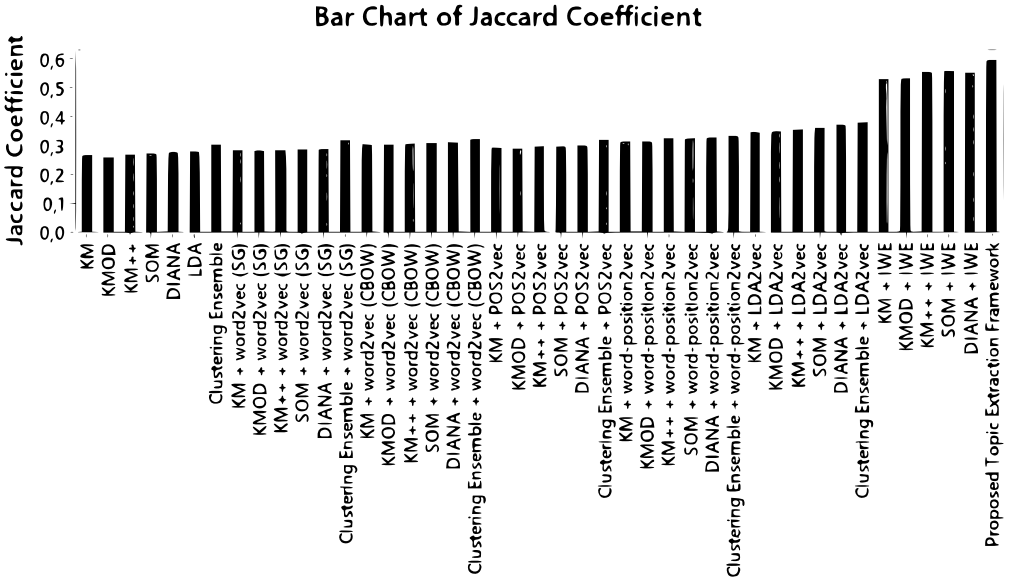


Fig. 1. Jaccard Coefficient metric for different combinations of embedding and clustering methods.

form of (sentiment, aspect). For example, a review of a restaurant could output (good, ambiance) or (bad, food).

For this finesse task a two-step procedure called JASen was developed. It consists of sentence embedding and classification via convolutional neural networks (CNNs). The former takes labeled data for embedding training and unlabeled data for further classification.

The results shown by the JASen model were compared with those of previously developed approaches. Values of aspect identification and sentiment polarity classification are presented in the tables 1 and 2 below ([3]).

Table 1. Quantitative evaluation on aspect identification (%).

| Methods | Restaurants | | | | Laptops | | | |
|---|---|---|---|---|---|---|---|---|
| | Accuracy | Precision | Recall | macro-F1 | Accuracy | Precision | Recall | macro-F1 |
| CosSim | 61.43 | 50.12 | 50.26 | 42.31 | 53.84 | 58.79 | 54.64 | 52.18 |
| W2VLDA | 70.75 | 58.82 | 57.44 | 51.40 | 64.94 | 67.78 | 65.79 | 63.44 |
| BERT | 72.98 | 58.20 | 74.63 | 55.72 | 67.52 | 68.26 | 67.29 | 65.45 |
| JASen | 83.83 | 64.73 | 72.95 | 66.28 | 71.01 | 69.55 | 71.31 | 69.69 |

Table 2. Quantitative evaluation on sentiment polarity classification (%).

| Methods | Restaurants | | | | Laptops | | | |
|---|---|---|---|---|---|---|---|---|
| | Accuracy | Precision | Recall | macro-F1 | Accuracy | Precision | Recall | macro-F1 |
| CosSim | 70.14 | 74.72 | 61.26 | 59.89 | 68.73 | 69.91 | 68.95 | 68.41 |
| W2VLDA | 74.32 | 75.66 | 70.52 | 67.23 | 71.06 | 71.62 | 71.37 | 71.22 |
| BERT | 77.48 | 77.62 | 73.95 | 73.82 | 69.71 | 70.10 | 70.26 | 70.08 |
| JASen | 81.96 | 82.85 | 78.11 | 79.44 | 74.59 | 74.69 | 74.65 | 74.59 |

In conclusion, the authors have proposed a model fine-tuned for joint aspect-sentiment extraction which outperformed earlier methods for this task.

However, more often the extraction of topics is enough. In contrast to the ensemble approach proposed in the publication [4], authors of the article [2] leverage different modified algorithms that compose BERTopic – a modern and efficient topic analysis tool.

Its defining feature is the ability to extract topics dynamically and to build topic popularity time series. This is allowed by using c-TF-IDF (class term frequency inverse document frequency) representations of topics:

$$\text{c\_\_TF\_\_IDF} = \underbrace{\frac{n_t}{\sum\limits_{k \in \text{class}} n_k}}_{\text{c\_\_TF}} \cdot \underbrace{\log\left[\frac{|D|}{|\{d_i \in D | t \in d_i\}|}\right]}_{\text{IDF}} \tag{4}$$

Here c-TF is defined by the number of instances of word $t$ divided by the number of words in the class, $|D|$ - number of documents in the dataset, $|\{d_i \in D | t \in d_i\}|$ – number of documents in dataset, containing the word $t$. First, BERTopic is fitted on the entire dataset to create a global view of topics. Using c-TF-IDF is efficient, as IDF is computed globally and only c-TF needs to be computed at each timestep.

In the publication, newly proposed BERTopic is compared to its predecessors: LDA, NMF, top2vec and CTM by means of topic coherence (TC) and topic diversity (TD) metrics on three datasets of news and twitter posts. According to the article, BERTopic has high coherence scores across all datasets. In terms of topic diversity, it is outperformed by CTM.

One of the major strengths of BERTopic is that it has the capability to operate in the multilingual mode. The default embedding model "all-MiniLM-L6-v2" shows both moderately high topic coherence and diversity across all supported languages in addition to it being lightweight.

More detailed comparison of BERTopic with other methods of topic analysis is presented in the publication [1]. The authors have chosen latent Dirichlet allocation (LDA), non-negative matrix factorization (NMF), top2vec and BERTopic in the task of topic analysis on the dataset of Twitter posts.

LDA is a generative probabilistic model for discrete data, which could be viewed as three-level hierarchical Bayesian clustering algorithm. Each document is represented as a mixture of topics with corresponding probabilities and each topic is a mixture over the collection of topic probabilities.

NMF is a decompositional method which works of TF-IDF transformed data by breaking down the input term-document matrix ($A$) into a pair of lower-ranking matrices:

- terms-topics ($W$) matrix containing basis vectors;
- topics-documents ($H$) matrix containing weights.

In NMF all elements in those matrices are non-negative so as to be interpretable.

Top2vec algorithm uses word embeddings via pretrained embedding models so that semantically close words have spatially close embedding vectors. Due to the sparsity of the vector space, a dimension reduction is performed before clustering. Commonly, hierarchical density-based spatial clustering of applications with noise (HDBSCAN) is used to identify dense regions in the reduced vector space of documents. As words that appear in several documents cannot be assigned to one document, they are recognized as noise. Thus, no lemmatization is needed beforehand.

BERTopic uses BERT pretrained embedders alongside with sentence-transformers for turning documents into vectors across more than 50 languages. Similarly to top2vec, BERTopic uses uniform manifold approximation and projection (UMAP) for dimension reduction and HDBSCAN for clustering. The principal difference between BERTopic and top2vec is that the latter utilizes c-TF-IDF metric instead of normal TF-IDF. Thus, importance of words in clusters and not in documents is taken into account. The usage of HDBSCAN eliminates the need for lemmatization.

Human interpretation of the extracted topics has shown that BERTopic and NMF perform the best among four tested methods. Despite that top2vec and BERTopic both use pretrained embedders for the

representation of documents in vector space, many topics extracted by top2vec contained several semantic concepts or intersected each other. LDA gave the least legible results of all methods.

The authors mentioned several drawbacks of BERTopic models such as them yielding large numbers of clusters which leads to the need in manual topic processing. Another mentioned disadvantage is that one document could potentially be assigned to just one topic, which often does not reflect reality. In spite of this, BERTopic still can be characterized as one of the most successful and universally applicable methods of topic analysis due to high model quality and efficiency. Thus, BERTopic was the topic analysis method of choice in the current research work.

### 2.2. Time series prediction

### 2.3. Prediction with exogenous variables

### 2.4. Impact of context information on consumption

## 3. Conclusion

## Acknowledgements

## References

[1] Egger, R., Yu, J., 2022. A topic modeling comparison between lda, nmf, top2vec, and bertopic to demystify twitter posts. Frontiers in Sociology 7.

[2] Grootendorst, M., 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. arXiv preprint arXiv:2203.05794 .

[3] Huang, J., Meng, Y., Guo, F., Ji, H., Han, J., 2020. Weakly-supervised aspect-based sentiment analysis via joint aspect-sentiment topic embedding. arXiv preprint arXiv:2010.06705 .

[4] Onan, A., 2019. Two-stage topic extraction model for bibliometric data analysis based on word embeddings and clustering. IEEE Access 7, 145614–145633.