Literature review

### Reviews analysis

An important source of knowledge about e-commerce customers are textual reviews. They can serve as a rich source of feedback for what in the shop or product is liked and what needs change. Also, in the textual reviews one can get to know customer’s opinions way better then using other types of feedback, for example 1-5 rating of a purchase. With these advantages, they come at the expense of increased complexity of such analysis. A big challenge is to extract meaningful information from this type of highly unstructured data.

Two most important types of text mining in text reviews is *sentiment prediction* and *topic mining* (in the context of reviews also often called *aspect mining*). Topic modeling is particularly challenging, as usually one does not have a annotated dataset with topics assigned to each text. That is why an unsupervised approach usually has to be used.

A go-to model for inferring the topic of a text is Latent Dirchlet Allocation (Blei, Ng, and Jordan 2003). The method is based on assumption, that each document is a mixture of a small number of topics. At the same time, each topic can be characterized by a distribution of words frequency.

Unfortunately, LDA approach was created with different purpose in mind. Typically reviews in the aspect of e-commerce are very short. Hong and Davison (2010) showed that LDA is not able to find informative topics in Twitter posts. These posts are bound by the rules of the platform to be shorter than 280 characters long. Possible reason that LDA does not cope well is that assumption about a document being mixture of topics is false. Short texts probably comprise of very small amount of topics, usually only one.

The drawbacks of LDA in setting of short texts were adressed by (Yin and Wang 2014) . They used Gibbs Sampling algorithm for the Dirichlet Multinomial Mixture model, which is improvement over typical LDA. The algorithm used by the authors is called Movie Group Process. Short introduction to this algorithm is included below.

Imagine a movie discussion group. There are k tables, and the goal is to assign students to tables according to their similar movie taste. There are 2 preference parameters set for each student:

1. Choose a table with students having similar movie taste. This is meant to introduce homogeneity of the clusters.
2. Choose a table with more students in this group. This rule is meant to improve completeness - so to the clusters have a reasonably high number of members.

The authors show that this algorithm provides superior performance to vanilla LDA not only when the texts are short, but also in general.

In recent years completely new approaches to Natural Language Processing emerged, thanks to improvement in the area of Neural Network algorithms. Two approaches are especially important as they serve as a baseline for the most recent findings in aspect (topic) recognition area. These two are word vector representations and attention mechanism. A short introduction of these two methods is presented in the section below.

In 2013, word2vec (Mikolov et al. 2013) was presented. The goal of this method is to learn a meaningful vector representation of each word in a corpus. Word2vec’s approach is to train a model that predicts all of the neighboring words for every occurrence of every word in an entire body of text (a corpus).

Intuitively, suppose that the model needs to learn embeddings for 3 words: “king,” “queen,” “orange.” The points in the embedding space for the first two words should lay in the proximity, while “orange” should be further. Word2vec approach is to look at the probability, that given word should be placed in particular place in the sentence, given the neighboring words. Suppose we have an incomplete sentence “XXX were usually very rich in the past.” Word2vec tries to predict what XXX should be. From the corpus it should understand, that “king” and “queen” are more probable than “orange,” that is why puts the embeddings closer.

Creating word embeddings usually serve as a preprocessing phase for next analysis steps, as with the data in numeric form one can use all tools that conventional data analysis has to offer, not being limited anymore by the complicated nature of textual data.

Another concept very helpful in the aspect recognition domain is attention mechanism (Chorowski et al. 2015). It is based on attention mechanism in psychology. When a human is trying to understand any content (visual, textual etc.) she is not using all content in the same extent, but only the relevant parts. For example, when a car driver is making a decision whether to cross an intersection, from all the visual signals that she obtains at the moment, the most important (and the only one looked at) is whether the light is red or green.

This concept can be very useful in the area of aspect prediction, as usually only couple of words from the whole sentence show the topic of it.

He et al. (2017) presented an Attention-based Aspect Extraction model. At first, words embedding using Word2Vec model is created. After that, for each text in the corpus, attention weight for each word is computed using neural network with an attention layer. Then, embedding of the whole sentence is created by computing an average for all words embedding. The words are weighted by their attention weights. Last step of the procedure is creating encoder-decoder model for learning sentence aspect embedding. The reconstruction of the sentence is the linear combination of aspect embeddings, and aspect embeddings are learned by mapping sentence embedding to a lower dimensional space.

Another work worth mentioning is by Tulkens and Cranenburgh (2020), who proposed a new type of Attention mechanism, meant especially for aspect recognition task. It’s advantage over the one presented by (He et al. 2017) is that instead of a complex neural network, a way simpler approach based on Radial Basis Function kernel is used. Another work presenting new attention mechanism is by Luo et al. (2019) - they use a use a Encoder-Decoder framework with an *Semene Attention* mechanism.

### Churn analysis

TODO:

* co to churn
* churn vs retention
* dlaczego się opłaca mieć lojalnych klientów

Although churn prediction is typically modeled as a classification problem, one should bear in mind the final goal of churn prediction, that is preventing the customer from leaving (Wai-Ho Au, Chan, and Xin Yao 2003). One strategy to do that is to target the most risky customers with some kind of marketing campaign. The ultimate goal for the model should be to assess the probability of customer leaving, and let the marketing experts decide what level of probability bothers the company, and how strong the reaction should be. For example, one could decide that for top 1% of the most risky customers the company should contact the customer in person by phone, for top 10% offer special discounts, while for top 30% - just send an encouraging email.

Because of that one should bear in mind that the models not outputting probabilities (for example vanilla SVM) are not the best choices. That is also why industry standard for churn prediction is logistic regression, as the outputs can be directly interpreted as probabilities, and also the customers can be ranked.

The most low-hanging fruit for the companies that want to start basing their business desicions on the data is usage of transaction-level data. That is because virtually every e-commerce shop is based on the the mechanism of user registration, and storing the client’s purchasing history is an industry standard. The data only about when the customer made purchases and how much did he pay are very easily translated into the framework of Recency-Frequency-Monetary value. Multiple works (Aleksandrova (2018), Yanfang and Chen (2017)) demonstrated that such data can serve as a good input to churn prediction machine learning model. In fact, most of the publications presented in this review is using RFM variables as one part of the dataset, while including more complex, engineered variables as the other part.

Besides the application in churn prediction domain, clustering the customer base into different segments based on RFM model can be very valuable just for business intelligence application and exploratory data analysis.

Because of the digital nature of e-commerce shopping, way more detailed and enriched data can be used in hope of finding more appropriate features. One of such features is per-session data - that is the information about how the user is navigating on the site. Yu et al. (2011) used the data avaliable in the data warehouse of e-commerce company to predict churn. They combined per-transaction data, per-session data and customer data using Extract-Transform-Load tools and manual feature engineering. Berger and Kompan (2019) tried to predict user churn on a per-session basis. The question they stated was “Will the customer unregister from the service during this session?” They used a very detailed per-session metadata like the day of the week, session number or number of purchases done up to the point.

After the first purchase of the customer in the e-commerce shop, their exact adresses can be inferred with high probability. Usually the delivery adress would be to the home of a customer, or in worse cases to other place that the customer visits (like workplace etc.). Zhao et al. (2005) used this kind of customer location data to enrich the dataset with basic spatial characteristics of the region, that is geographic situation and demographic variables.

The content of the web is very often avaliable in unstructured, textual form. Online retailers very often give their customers to write reviews about the purchase. Although such information is very hard to incorporate into churn prediction model, it can serve as a very rich source of insight. De Caigny et al. (2020) showed how such data can be incorporated into churn prediction modeling to obtain superior results. They have used text embedding approach as a feature extraction method.

Other branch of customer behaviour analysis is concentrated on declaration of desire to repurchase in particular shop. Compared to the actual analysis of sales data, this approach has an advantage that the customer can have a very good experience in the shop, but not do the repurchase from other reasons. This information is still very valuable to the shop. On the other hand, questionaire data has usually less quality as the customers can lie, an also response rate is usually very low.

Suryadi (2020) analysed responses to such questionaires and tried to predict, whether preprocessed textual reviews can serve as explanatory variables. After using tf-idf transformation they found out that they can.

Methods description

### Methods - text reviews

Olist e-commerce store is operating only in Brazil. That is why most of the reviews are written in Portuguese. I have used Google Translate API to change the language of them to English. This is to facilitate not only understanding the reviews, but also the NLP tools avaliable for english language are more up-to-date and advanced.

3 algorithms for topic modeling that I have tested were:

1. Latent Dirchlet Allocation - because it is a go-to standard for topic recognition
2. Gibbs Sampling algorithm for the Dirichlet Multinomial Mixture - as this method is an improvement over LDA, meant especially for short texts. This is the case from this study, as most of the reviews are just couple of words long
3. Attention-Based Aspect Extraction - this method is also meant for short texts, and at the same time it uses the most modern, state-of-the-art NLP techniques.

To improve the results in all 3 methods I have removed stopwords and punctuation from the text. Also, to limit the number of words in the vocabulary I have used lemmatization technique. This was done by WordNet lemmatizer. To further improve the results, I have used Part-of-Speech tagger, and passed the tags of words to the lemmatizer. This way the algorithm can change the form of the word on a more informed basis, and thus apply lemmatization to more words.

Second step of the preprocessing was converting lemmatized reviews into vector format. In case of LDA, count-vectorizing approach was applied. the words which appeared in less than 0.1% of reviews were dropped from the dataset. In the case of Gibbs Sampling the same preprocessing is done internally by the function.

In both of these cases after vectorization one should obtain a matrix with n rows and k columns, where n is number of observations in the original dataset, while k - size of the vocabulary.

Very different preprocessing was required in the case of Attention-Based Aspect Extraction. The neural network requires simply lemmatized reviews in textual format as the output. Then, one of the layers of the network is meant to embed the currently preprocessed word. These embeddings are not learnt during the network training, they should be trained beforehand instead. The authors of the paper propose Word2vec technique for learning embeddings. Following their guidelines I have used this method, setting the dimensionality of the vector space to 200. I have also applied the word window of 10.

After applying word2vec on this dataset, I have obtained the matrix with m rows and 200 columns, where m stands for number of words in the dataset, and 200 is the dimensionality of the vector space chosen as a hyperparameter. One should bear in mind that count-vectorization works on review level, while word2vec - on words level. This means that after applying word2vec model to 1 review, one would obtain the same number of outputs as the number of words in the review. This is why it is impossible to use word2vec to preprocess the dataset and then use LDA or Gibbs sampling without some way to convert couple of vectors into one.

Unfortunately, the evaluation of topic extraction is a hard task. The only reasonable one is human inspection. That is why after running every model I have verified the obtained topic for coherency and distinctiveness.

TODO: Opis wyboru parametrów modeli z poniższej listy

1. LDA
   1. remove stopwords and punctuation
   2. use wordnet lemmatizer (Fellbaum 1998) with POS tagging (NLTK)
   3. use count-vectorize approach with removing words which appear in less than 0.1% reviews
   4. use LDA with varying number of components 3, 5, 10, 15
   5. manually inspect inferred topics
2. Gibbs Sampling algorithm for the Dirichlet Multinomial Mixture (movie group process)
   1. same as in LDA but this one does not require vectorizing the dataset
   2. 30 topics
   3. Citing python gsdmm package documentation:
   * “Alpha controls the probability that a student will join a table that is currently empty. When alpha is 0, no one will join an empty table. Beta controls the student’s affinity for other students with similar interests. A low beta means that students desire to sit with students of similar interests. A high beta means they are less concerned with affinity and are more influenced by the popularity of a table”
   1. checking every combination of alpha, beta chosen from the values: [0.01, 0.1, 0.5, 0.9]
   2. manual inspection
3. Attention-Based Aspect Extraction (He et al. 2017)
   1. Based on the implementation by authors of the article
   2. remove stopwords and punctuation
   3. use wordnet lemmatizer (Fellbaum 1998)
   4. converting to vectorized form using Word2Vect model with final dimensionality of 200 and window of 10
   5. Training attention model
   6. manual assessment of reviews topics

For the analysis of textual reviews I have used python programming language.

1. python (Van Rossum and Drake Jr 1995)
   1. NLTK (Bird, Klein, and Loper 2009), spacy (Honnibal et al. 2020) itd

Luźne opisy paperów:

<https://link.springer.com/chapter/10.1007/978-981-32-9563-6_11> (Jheng and Luo 2019) - retention prediction przez CNN

<https://www.sciencedirect.com/science/article/pii/S0019850116301651?casa_token=YCUcElM8k_EAAAAA:-qJeOGXh7u2pQlqj-eyAo9k-eLgbc-m31QsDURsmpD2CEIyqtUzAjYGXUwkQRR4T0MrtkIbeWtaG>

* sam paper średni ale ma dużo referencji

<https://sci-hub.se/https://ieeexplore.ieee.org/abstract/document/7538581>

* fancy metody imbalance
* używają datasetu takiego jak ja
* mega dobra dokładność
* ALE najpierw robią upsampling a później oceniają performance na CV - błąd! u mnie to różnica pommiędzy 0.62 a 0.78 AUC

<http://flr-journal.org/index.php/mse/article/view/10816/11113>

* wykorzystują social network userów

<https://link.springer.com/article/10.1007/s10660-019-09383-2>

* wykorzystują deep learning do imbalance - chyba lepiej będzie pasować do innej sekcji

<https://www.emerald.com/insight/content/doi/10.1108/17515631011063767/full/html>

dużo references

A study on factors affecting the purchasing process of online shopping: a survey in China & Japan

na podstawie kwestionariusza ocena satysfakcji

<https://d1wqtxts1xzle7.cloudfront.net/62198454/key-paper20200225-3623-15suux9.pdf?1582687757=&response-content-disposition=inline%3B+filename%3DWhat_Effects_Repurchase_Intention_of_Onl.pdf&Expires=1617187287&Signature=LfylLp7R2PXNPLFtVyCNdj~e4FhDBUz04-T152E7FSsNHjnqclWeFnnKf9C2fJskRN2q~sRx~CsXCbeuhn0zcrktL0lj8oN8GUxrWXpavIz1UaQuO~ayrylqfAH2XgIdwhDe~8FOoMNP9ZzaNz6lqYuy6DYaBNhP6G7N3sUo2spQ187dGOgRHgGafoS3Z7HZ2AgEUjgs1ldOsU1E7FXrP1delDpO7QYarp9h1euOUM6vCWCxlsDZYnRF6A-PIuQlgyP8QOyzMo2d487sDw0Jepwjrd69ocCrSMsi7dmu56Z00CUoXaUA3b~C9vyQrfYI9T1hzMcJYfQYri4lUWgblQ__&Key-Pair-Id=APKAJLOHF5GGSLRBV4ZA>

ocena repurchase probability na podstawie kwestionariusza

<https://www.tandfonline.com/doi/abs/10.1080/08874417.2011.11645518?casa_token=33mj-Wcpw8IAAAAA%3AQUNps1MzrKJLyKf_c0Vl6gRIzoqI8wU3PbavgeTiyiJlQxwpHMi3JLUMmmGr7ZX0C2uqsrTT-TBYIw8&>

znowu kwestionariusz ale z modelem

<https://link.springer.com/article/10.1007/s10660-015-9207-2>

* predicting repurchase intention
* na podstawie kwestionariusza
* ale prediction z fancy metodami

<https://ieeexplore.ieee.org/abstract/document/9325646> (Suryadi 2020)

* repurchase jako kwestionariusz tak/nie
* ale wykorzystują predykcję na podstawie reviews

(Ganesh, Arnold, and Reynolds 2000) definicja churn

A particularly prominent forecasting application in CRM is customer churn prediction (CCP), which is defined as a method of identifying customers who show a high inclination to abandon the company

<https://www.sciencedirect.com/science/article/pii/S0169207019301499?casa_token=kopLN0D45dwAAAAA:pARTYFQ1-0aho11qk4RpZdFdBIb1S-cJVHPb1iaggq41zU7pI-heeNpG9uK5cGThM7IWfFAkeGqU> (De Caigny et al. 2020)

-ładnie opisane profity z posiadania lojalnych customerów - dobry paper, dużo odniesień i wykorzystanie textual data

<https://www.sciencedirect.com/science/article/pii/S0957417410006779?casa_token=0C1SeJigqT8AAAAA:GCfX81AUr9p3ZfrqwTPCb23r4Slx6YijCvIOJE5xTcrxgl1nge7gjwvQnCo4c_r5fp1zaSigKjve> (Yu et al. 2011)

* Jest o prawdziwym churnie a nie o retention
* Jest złożona baza danych
* minimalny wstęp o churn prediction

<https://ieeexplore.ieee.org/abstract/document/8627369> (Berger and Kompan 2019)

* używają danych o sesji w przeglądarce

<https://cursa.ihmc.us/rid=1MYWPTN4Z-BBB2D6-30SB/Zhao_Churn_Prediction_SVM.pdf>

(Zhao et al. 2005) - Używa danych demograficznych - jest o churn

<https://ieeexplore.ieee.org/abstract/document/8284914> Yanfang and Chen (2017) - prawie nic nie ma ciekawego, tylko jako case

<https://ieeexplore.ieee.org/abstract/document/1255389> (Wai-Ho Au, Chan, and Xin Yao 2003)

* Jest o tym że nie chodzi o predykcję tylko o ranking

##### References

(Tulkens and Cranenburgh 2020) CAT attention (Chorowski et al. 2015) word2vec (Mikolov et al. 2013) Main: (He et al. 2017) drawbacks of LDA (Hong and Davison 2010)

Aleksandrova, Yanka. 2018. “Application of Machine Learning for Churn Prediction Based on Transactional Data (RFM Analysis).” In *18 International Multidisciplinary Scientific Geoconference SGEM 2018: Conference Proceedings*, 18:125–32. 2.1.

Berger, P., and M. Kompan. 2019. “User Modeling for Churn Prediction in e-Commerce.” *IEEE Intelligent Systems* 34 (2): 44–52. <https://doi.org/10.1109/MIS.2019.2895788>.

Blei, David M, Andrew Y Ng, and Michael I Jordan. 2003. “Latent Dirichlet Allocation.” *The Journal of Machine Learning Research* 3: 993–1022.

Chorowski, Jan, Dzmitry Bahdanau, Dmitriy Serdyuk, Kyunghyun Cho, and Yoshua Bengio. 2015. “Attention-Based Models for Speech Recognition.” *arXiv Preprint arXiv:1506.07503*.

De Caigny, Arno, Kristof Coussement, Koen W. De Bock, and Stefan Lessmann. 2020. “Incorporating Textual Information in Customer Churn Prediction Models Based on a Convolutional Neural Network.” *International Journal of Forecasting* 36 (4): 1563–78. https://doi.org/<https://doi.org/10.1016/j.ijforecast.2019.03.029>.

Ganesh, Jaishankar, Mark J Arnold, and Kristy E Reynolds. 2000. “Understanding the Customer Base of Service Providers: An Examination of the Differences Between Switchers and Stayers.” *Journal of Marketing* 64 (3): 65–87.

He, Ruidan, Wee Sun Lee, Hwee Tou Ng, and Daniel Dahlmeier. 2017. “An Unsupervised Neural Attention Model for Aspect Extraction.” In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vancouver, Canada: Association for Computational Linguistics.

Hong, Liangjie, and Brian D Davison. 2010. “Empirical Study of Topic Modeling in Twitter.” In *Proceedings of the First Workshop on Social Media Analytics*, 80–88.

Jheng, Yang Ruei, and Zhi Chao Luo. 2019. “Customer Retention Prediction with CNN.” In *Data Mining and Big Data: 4th International Conference, DMBD 2019, Chiang Mai, Thailand, July 26–30, 2019, Proceedings*, 1071:104. Springer.

Luo, Ling, Xiang Ao, Yan Song, Jinyao Li, Xiaopeng Yang, Qing He, and Dong Yu. 2019. “Unsupervised Neural Aspect Extraction with Sememes.” In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, 5123–29. International Joint Conferences on Artificial Intelligence Organization. <https://doi.org/10.24963/ijcai.2019/712>.

Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. “Efficient Estimation of Word Representations in Vector Space.” *arXiv Preprint arXiv:1301.3781*.

Suryadi, D. 2020. “Predicting Repurchase Intention Using Textual Features of Online Customer Reviews.” In *2020 International Conference on Data Analytics for Business and Industry: Way Towards a Sustainable Economy (ICDABI)*, 1–6. <https://doi.org/10.1109/ICDABI51230.2020.9325646>.

Tulkens, Stéphan, and Andreas van Cranenburgh. 2020. “Embarrassingly Simple Unsupervised Aspect Extraction.” In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 3182–87. Online: Association for Computational Linguistics. <https://doi.org/10.18653/v1/2020.acl-main.290>.

Wai-Ho Au, K. C. C. Chan, and Xin Yao. 2003. “A Novel Evolutionary Data Mining Algorithm with Applications to Churn Prediction.” *IEEE Transactions on Evolutionary Computation* 7 (6): 532–45. <https://doi.org/10.1109/TEVC.2003.819264>.

Yanfang, Q., and L. Chen. 2017. “Research on e-Commerce User Churn Prediction Based on Logistic Regression.” In *2017 IEEE 2nd Information Technology, Networking, Electronic and Automation Control Conference (ITNEC)*, 87–91. <https://doi.org/10.1109/ITNEC.2017.8284914>.

Yin, Jianhua, and Jianyong Wang. 2014. “A Dirichlet Multinomial Mixture Model-Based Approach for Short Text Clustering.” In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 233–42. KDD ’14. New York, NY, USA: Association for Computing Machinery. <https://doi.org/10.1145/2623330.2623715>.

Yu, Xiaobing, Shunsheng Guo, Jun Guo, and Xiaorong Huang. 2011. “An Extended Support Vector Machine Forecasting Framework for Customer Churn in e-Commerce.” *Expert Systems with Applications* 38 (3): 1425–30.

Zhao, Yu, Bing Li, Xiu Li, Wenhuang Liu, and Shouju Ren. 2005. “Customer Churn Prediction Using Improved One-Class Support Vector Machine.” In *International Conference on Advanced Data Mining and Applications*, 300–306. Springer.

Bird, Steven, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. " O’Reilly Media, Inc.".

Fellbaum, Christiane. 1998. *WordNet: An Electronic Lexical Database*. Bradford Books.

He, Ruidan, Wee Sun Lee, Hwee Tou Ng, and Daniel Dahlmeier. 2017. “An Unsupervised Neural Attention Model for Aspect Extraction.” In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vancouver, Canada: Association for Computational Linguistics.

Honnibal, Matthew, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. *spaCy: Industrial-strength Natural Language Processing in Python*. Zenodo. <https://doi.org/10.5281/zenodo.1212303>.

Van Rossum, Guido, and Fred L Drake Jr. 1995. *Python Tutorial*. Centrum voor Wiskunde en Informatica Amsterdam, The Netherlands.