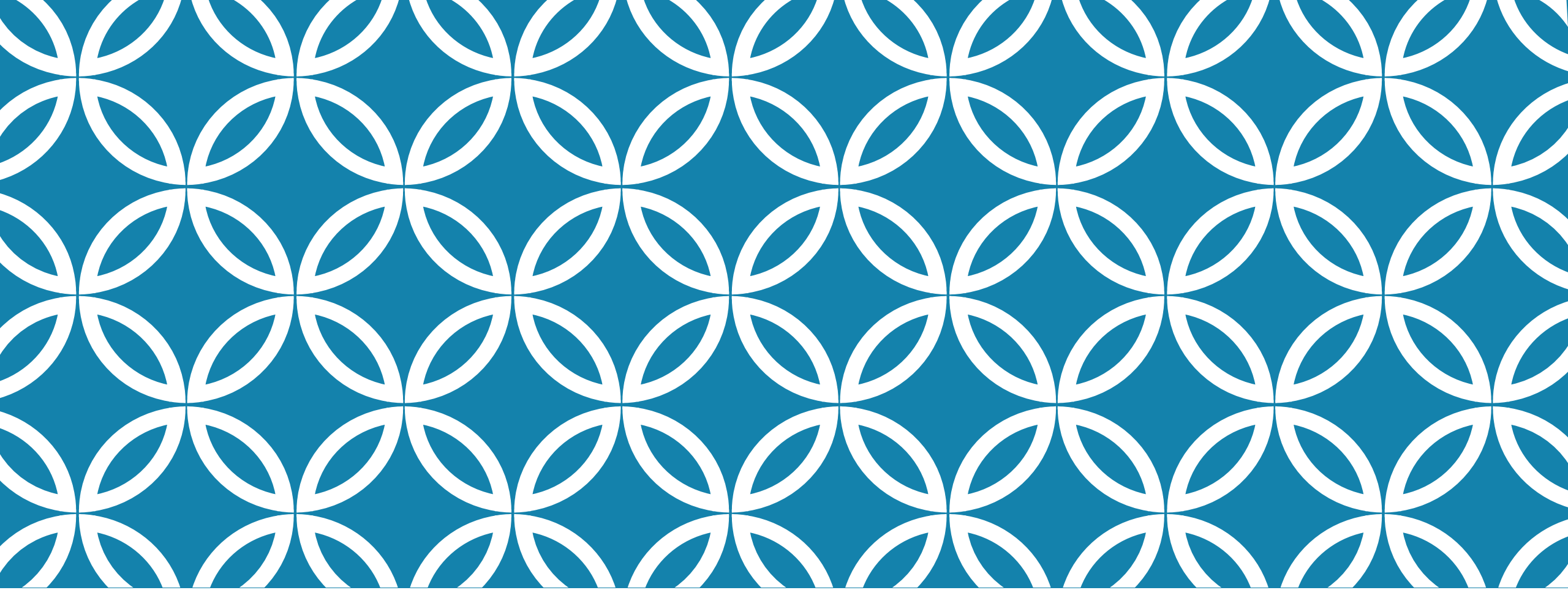

BATTLE OF THE NEIGHBORHOOD

Using data mining techniques to form a marketing strategy for a company

TABLE OF CONTENT

1. INTRODUCTION/BUSINESS PROBLEM

The purpose of the thesis is to develop a methodology for conducting a marketing research based on data mining using the example of a model for processing reviews of restaurants that determine the emotional color of a review. The data were taken from a Foursquare API. It contains reviews of various establishments from users. In addition to reviews, detailed information about each organization is also provided, including characteristics, as well as information about users.

The object of the research is the company of the restaurant business. The subject is the process of forming a marketing strategy for restaurant development, taking into account the information contained in customer reviews.

The following tasks were set:

- Search for practical examples of using feedback analysis using machine and in-depth training;
- The study of methods for the intellectual analysis of texts;
- Data preparation for model building;
- Building models;
- Analysis of the results and selection of the best model.

2. DATA

Most data mining methods were designed to work with indicators that describe consumers in one way or another: income, click rate, social group and place of residence, gender and age, education, purchase history, and so on. To collect and store all these data, you need the right infrastructure, the creation of which requires qualified specialists. Moreover, it is desirable to have a part of the data over time to build mathematical models, and, therefore, it takes time to obtain them. But what to do if the company does not have suitable infrastructure or it entered the market not so long ago. You can collect focus groups and organize their survey, you can run beta testing, but sometimes the most suitable tool is to study reviews. This does not require long data collection, which is undoubtedly an advantage.

2. DATA

Foursquare Location Data

The Foursquare location data will be used to explore Manhattan and find the tips and comments to specific restaurant "3 Guys Restaurant".

We use Foursquare API to get tips that are in "3 Guys Restaurant" in Manhattan within a radius of 500 meters, and transform the data into dataframe with 7 columns: ['Neighborhood', 'Neighborhood Latitude', 'Neighborhood Longitude', 'Venue', 'Venue Latitude', 'Venue Longitude', 'Venue Category']

Analyze neighborhood by one Restaurant, grouping rows by venue and taking the mean of the frequency of occurrence of each 'Venue Category', and then run Word2Vec and convolutional neural networks.

2. DATA

```
dataframe_filtered
```

[6]:

	name	categories	address	cc	city	country	crossStreet	distance	formatted
0	3 Guys Restaurant	Diner	49 E 96th St	US	New York	United States	Madison Ave	570	[49 E (Madi New
1	Hanratty's Restaurant	Food	1410 Madison Ave	US	New York	United States	NaN	551	[1410 Madi New 10029,
2	Polonia Restaurant	Food	1398 Madison Ave	US	New York	United States	NaN	573	[1398 Ave, New 10029,

Visualize the Restaurants that are nearby

[7]:
```
dataframe_filtered.name
```

[7]:
```
0        3 Guys Restaurant
1    Hanratty's Restaurant
2       Polonia Restaurant
Name: name, dtype: object
```

3. METHODOLOGY

In order to train the model, the marked data is required. If we want to learn how to differ a positive feedback from a negative one, first, we need to assign a positive or negative mood tag to each feedback from the training set. Considering that in order to achieve a sufficient generalizing ability, the model must be trained on a large number of reviews, there are two ways to solve this problem. The first one is to use data from reference services, such as Foursquare, in which users, in addition to reviews, give marks to institutions. In the above service, there is API to access the necessary information. The second way is to hire assessors who will independently mark up the data. This method is more costly and time consuming, but it has to be used if you need to analyze specific areas for which there are no special services, as in the first case.

4. RESULTS

Word2Vec and convolutional neural networks

Now create a more complex model. Just as before, we will remove the excess from the reviews and make a lemmatization. We will only include in the model those words that occur more often than 30 times, the window size is 10, and each word will be represented by a dimension vector of 300. To check whether we got a good idea or not, choose several words and look for them gravest in vector space.

4. RESULTS


highly recommend first time
one favorite
great service
food service
great place
place great
next time
restaurant pittsburgh
love place
great food
great atmosphere
best pizza
food delicious
wait staff
food amazing
food great
delicious food
dining experience
favorite restaurant
last night
south side
make sure
ice cream
staff friendly
one best
best ever
happy hour
service great
every time
top notch

5. DISCUSSION

Application of the obtained models for the formation of a marketing strategy.

On a concrete example, we will show the applicability of the proposed models for marketing research. We will consider the city of New York. The first thing worth noting is that there are fewer reviews, even though there are almost 1,000 restaurants in New York. Choose positive and negative reviews using our model for "3 Guys Restaurant".

This is an old and famous Restaurant, which is positioned as a place where you can try the classic Canadian recipe for smoked meat. Most of its visitors are tourists. Therefore, it is important for an institution to maintain a reputation on the Internet. In total for this restaurant on the Foursquare website there are 16 reviews, and the average restaurant rating is 4 stars. Let's see what can be fixed to increase the rating to 5.

The accuracy of the model for this institution was 84%. Create word clouds for predicted and actual negative comments.

In figures below, the same main complaints of visitors can be distinguished: • Bad service • Long time in line: 30-45 minutes • Dry meat • Tight room In order to take advantage of thematic modeling, more reviews are needed. Therefore, we will continue to use Word2Vec to find out the preferences of visitors.

5. DISCUSSION

5. DISCUSSION

For this restaurant, the share of dissatisfied consumers, rated restaurants in 1.2 and 3 stars is about 23%. Thus, almost every fourth was dissatisfied with the visit, which, of course, adversely affects it's reputation. On the other hand, this means that there is potential for growth. It was found in the work that an increase in the restaurant rating by one star leads to an increase in income by 5–9%, depending on the particular restaurant.

6. CONCLUSION

In the course of this work, the main areas of application of text data processing methods in business were examined, and specific examples of their use in marketing were found. Various methods of processing texts and determining their tonality were tried. The best result was shown by direct propagation neural networks, to which the input was a bag of words with bigrams with TF-IDF weighting and a dictionary size of 10,000 words, as well as recurrent neural networks with GRU neurons. It was also found that the assessment on a five-point scale is quite subjective and each person can evaluate the same text differently. In the experiment, the neural network worked even more accurately than the survey participants.

6. CONCLUSION

The results of the study allow us to conclude that using machine learning, it is possible to achieve acceptable accuracy for the task of recognizing the key of the utterance, which allows real-time monitoring of user reactions to the products and activities of the company. Data for training models can be obtained using the API of large recommendation services. For institutions in Russia in particular, the Foursquare API can be used. In the future, for a broader picture, you can take comments and posts of users in social networks. You can use thematic modeling, word clouds, and Word2Vec to get an idea of what exactly users like in some area, and what causes them irritation. The proposed tools allow you to partially automate this process and learn from the opinion of a larger number of clients than regular surveys. But given the trends in the behavior of users of social networks, for future research, it is of interest to analyze the tonality of photos and videos of users, as well as short text messages related to these photos and videos. So, according to the results of research in 2017, the share of video content from all information in the network will be 74%, tweets with pictures are shared 1.5 times more often than without pictures, Facebook posts with pictures collect 2.3 times more, in December 2016 Instagram, the platform with the most pictures and short videos, announced a user base growth of 100 million over six months, and at the moment more than 700 million people use the service. Thus, much more information can be obtained by analyzing not only texts, but also photo and video content.