САМАРСКИЙ УНИВЕРСИТЕТ
SAMARA UNIVERSITY

Social Climate Analysis based on Open Data

Oleg Golovnin[1], Irina Dubinina[2], Anton Ivaschenko[3], Arkadiy Krivosheev[2], Pavel Sitnikov[4]

[1]Samara University, [2]Samara State Technical University, [3]Samara State Medical University, [4]Open Code SEC

Introduction

The paper proposes an approach to identifying problematic issues according to open sources for socially oriented topics. The approach is programmatically implemented on the basis of the Integrated Monitoring Digital Platform, which makes it possible to reduce the load on the analyst-operator, who analyzes the key problems of the region, by significantly reducing the amount of information viewed.

Methodology of Integrated Monitoring

Identification of problematic issues on socially oriented topics is carried out on the basis of the analysis of publications in open thematic groups and on public pages of users in social networks.

Stage 1. At the first stage, the initial data associated with the given analyzed administrative region is collected.

Stage 2. collected data is cleared of spam, for which a classifier is used that classifies the post as “spam” or “not spam”.

Stage 3. Search messages that contain occurrences of keywords.

Stage 4. The sentiment analysis of each message from the obtained filtered sample using a model based on BERT

Stage 5. The general topics and directions of publications are determined. Topic clustering is done as follows

1) Between user publications based on embeddings, a cosine measure of proximity between vectors A and B is considered

$$cosine\_measure = 1 - \frac{A \cdot B}{\|A\|\|B\|} = 1 - \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}}$$

2) If the cosine measure is less than 0.25, then user publications are defined as similar in meaning or value and are combined into one cluster;

3) After merging into clusters, an iterative enumeration of pairs of clusters is performed, and if more than half of the publications in the smaller cluster belong to the larger cluster, then the smaller cluster is merged with the larger cluster;

4) Active clusters are those whose size is at least $l = \max(3, L^{0,25})$, where L is the number of posts after spam checks;

5) Within each cluster, the number of bigrams (phrases) is counted, taking into account the syntax in sentences; N=3 most frequent digrams becomes the name of the cluster.

6) For each cluster, the sentiment level is estimated $T = \max(count(T_n))$, $n = 1..5$, where n corresponds to the sentiment category number of the text.

This approach makes it possible to determine the clusters, their names and tonality, with which the decision makers will further work. Thus, negative clusters are considered problematic issues of the region, and positive clusters are considered the most significant achievements of the region, neutral clusters do not contain acute social problems.

Software implementation

The software implementation of the proposed approach is based on the Integrated Monitoring Digital Platform as a software module. The software module code is written in Python 3.8 using the Django framework. The web interfaces are written in Vue JS. PostgreSQL 12.6 is used as a DBMS.

A diagram of the components of the Integrated Monitoring Digital Platform is shown in Fig. 1.

To classify social network publications for spam, a deep neural network with the following architecture is used: an input layer of 768 neurons, two hidden layers consisting of 256 and 128 neurons, leaky_relu with a leak parameter of 0.01 is used as an activation function for these layers, and an output layer of one neuron with a Sigmoid activation function.

Hits from social network users are collected into hit clusters using the built-in Python tools, with the exception of the algorithm for creating a name for a hit cluster. As the name for the cluster, a phrase is chosen that is most often found among the publications of users that make up the cluster.

Fig. 2 shows the software implementation of viewing clusters of citizens' appeals in the form of a widget developed in the interface of the Integrated Monitoring Digital Platform.

Clusters with a negative assessment of the sentiment of messages are marked as important problems of the region and have a priority status in the ranking of clusters. Based on the information received, decision makers can not only more effectively provide targeted assistance to those who apply, but also analyze problems in the region.

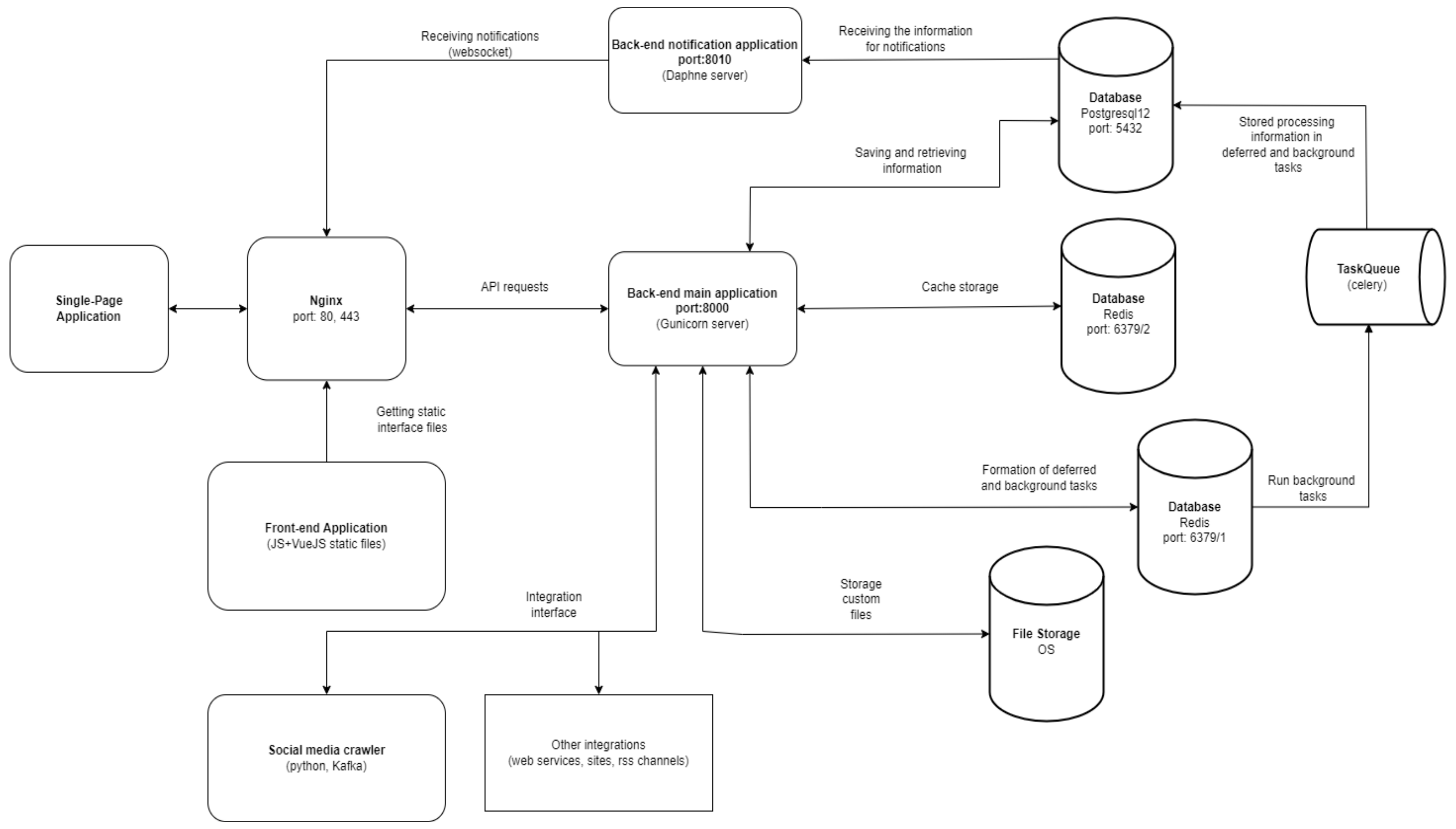
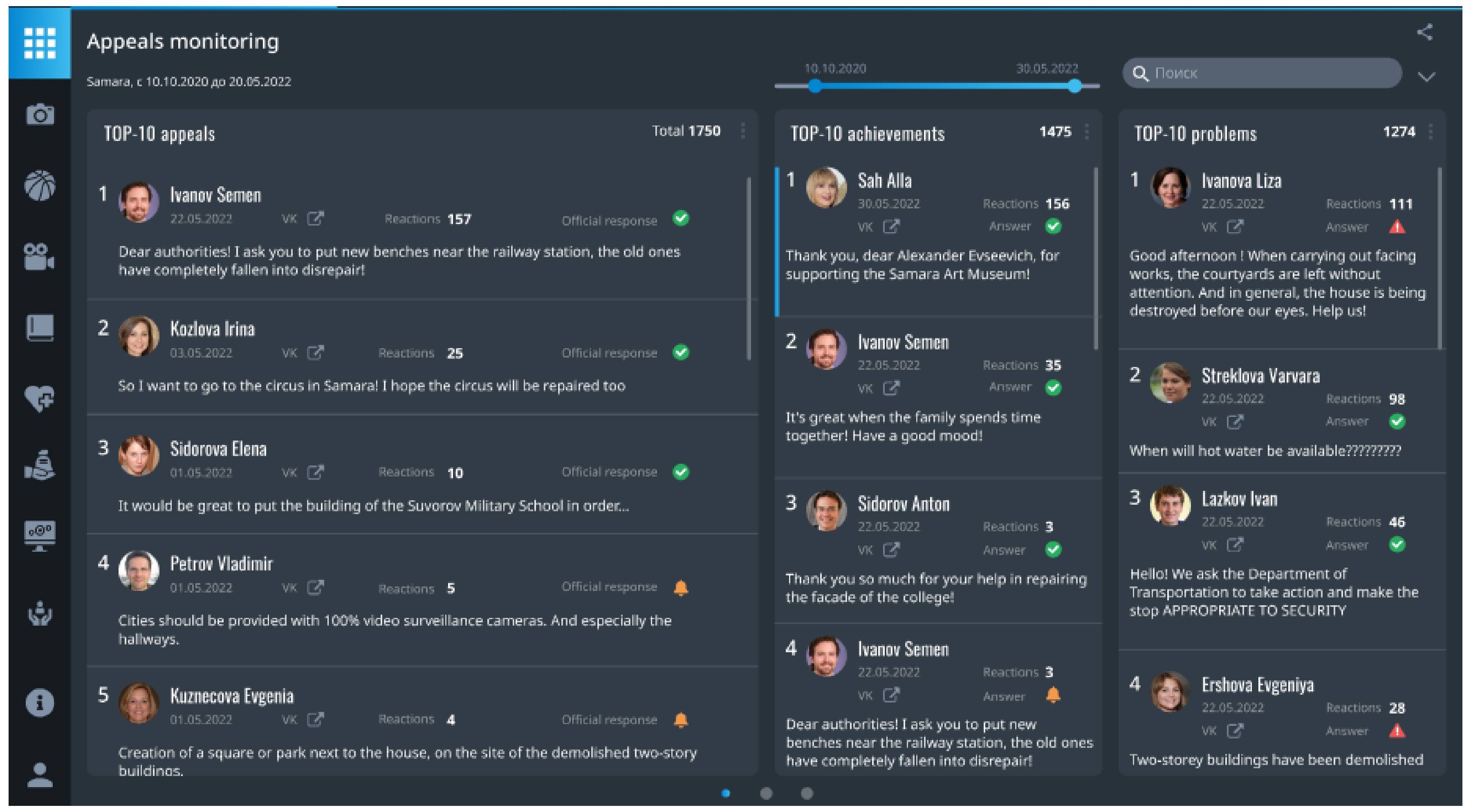


Fig. 1. Integrated Monitoring Digital Platform components



Samara, с 10.10.2020 до 20.05.2022

10.10.2020 30.05.2022

TOP-10 appeals Total 1750

1 Ivanov Semen 22.05.2022 VK Reactions 157 Official response ✓
Dear authorities! I ask you to put new benches near the railway station, the old ones have completely fallen into disrepair!

2 Kozlova Irina 03.05.2022 VK Reactions 25 Official response ✓
So I want to go to the circus in Samara! I hope the circus will be repaired too

3 Sidorova Elena 01.05.2022 VK Reactions 10 Official response ✓
It would be great to put the building of the Suvorov Military School in order...

4 Petrov Vladimir 01.05.2022 VK Reactions 5 Official response 🔔
Cities should be provided with 100% video surveillance cameras. And especially the hallways.

5 Kuznecova Evgenia 01.05.2022 VK Reactions 4 Official response 🔔
Creation of a square or park next to the house, on the site of the demolished two-story buildings.

TOP-10 achievements 1475

1 Sah Alla 30.05.2022 Reactions 156 VK Answer ✓
Thank you, dear Alexander Evseevich, for supporting the Samara Art Museum!

2 Ivanov Semen 22.05.2022 Reactions 35 VK Answer ✓
It's great when the family spends time together! Have a good mood!

3 Sidorov Anton 22.05.2022 Reactions 3 VK Answer ✓
Thank you so much for your help in repairing the facade of the college!

4 Ivanov Semen 22.05.2022 Reactions 3 VK Answer 🔔
Dear authorities! I ask you to put new benches near the railway station, the old ones have completely fallen into disrepair!

TOP-10 problems 1274

1 Ivanova Liza 22.05.2022 Reactions 111 VK Answer ⚠
Good afternoon ! When carrying out facing works, the courtyards are left without attention. And in general, the house is being destroyed before our eyes. Help us!

2 Streklova Varvara 22.05.2022 Reactions 98 VK Answer ✓
When will hot water be available?????????

3 Lazkov Ivan 22.05.2022 Reactions 46 VK Answer ✓
Hello! We ask the Department of Transportation to take action and make the stop APPROPRIATE TO SECURITY

4 Ershova Evgeniya 22.05.2022 Reactions 28 VK Answer ⚠
Two-storey buildings have been demolished

Fig. 2. Results of data processing in Integrated Monitoring Digital Platform

Results

For 3 months in the analyzed region, about 700 thousand posts were detected, of which 480 thousand posts were classified as spam. Of the remaining 220 thousand posts, only 50 thousand posts are socially significant, of these 50 thousand posts, 4 thousand posts are positive, and 18 thousand posts are negative. As a result of the analysis of positive posts, 9 clusters were identified, of which 7 were marked by the analyst as targeted – containing up-to-date information. In negative posts, 45 clusters were identified, of which 21 were marked by the analyst as targeted.

This approach makes it possible to determine the clusters, their names and tonality, with which the decision makers will further work. Thus, negative clusters are considered problematic issues of the region, and positive clusters are considered the most significant achievements of the region, neutral clusters do not contain acute social problems.