The Program for Public Mood Monitoring through Twitter Content in Russia

S.I. Smetanin <sismetanin@gmail.com>
National Research University Higher School of Economics,
20, Myasnitskaya st., Moscow, 101000 Russia

Abstract. With the popularization of social media, a vast amount of textual content with additional geo-located and time-stamped information is directly generated by human every day. Both tweet meaning and extended message information can be analyzed in a purpose of exploration of public mood variations within a certain time periods. This paper aims at describing the development of the program for public mood monitoring based on sentiment analysis of Twitter content in Russian. Machine learning (naive Bayes classifier) and natural language processing techniques were used for the program implementation. As a result, the client-server program was implemented, where the server-side application collects tweets via Twitter API and analyses tweets using naive Bayes classifier, and the client-side web application visualizes the public mood using Google Charts libraries. The mood visualization consists of the Russian mood geo chart, the mood changes plot through the day, and the mood changes plot through the week. Cloud computing services were used in this program in two cases. Firstly, the program was deployed on Google App Engine, which allows completely abstracts away infrastructure, so the server administration is not required. Secondly, the data is stored in Google Cloud Datastore, that is, the highly-scalable NoSQL document database, which is fully integrated with Google App Engine.

Keywords: sentiment analysis; public mood; mood patterns; twitter; social media

DOI: 10.15514/ISPRAS-2017-29(4)-22

For citation: Smetanin S.I. The Program for Public Mood Monitoring through Twitter Content in Russia. *Trudy ISP RAN/Proc. ISP RAS*, vol. 29, issue 4, 2017, pp. 315-324. DOI: 10.15514/ISPRAS-2017-29(4)-22

1. Introduction

With the popularization of social media, particularly the micro-blogging website Twitter, a vast amount of content is directly generated by people every day. In addition to textual information, which seems to have affective component, Twitter messages are also time-stamped and geo-located. Consequently, both tweets meaning and extended information about a message can be analyzed in a purpose of

scientific studies in general and in the exploration of public mood variations particularly.

In data mining, the usage of social media to analyze and predict political events is becoming more popular in recent times. During the Brexit referendum in the United Kingdom, the researchers consider changes to the public mood within the contents of Twitter [13]. They measure the appearance of positive and negative affect in various geographic regions of the United Kingdom, at hourly intervals. According to the results, there are three key times in the period leading up to and including the EU referendum, each of which was characterized by an increase in negative affect with a corresponding loss of positive affect.

The paper [12] describes an empirical study of Relationship between Twitter mood and stock market from an Indian context. Using Twitter as a source of the news, the authors have extracted the polarity of messages and have found a significant correlation with stock market movement measured in the major stock indices of India. In addition, the correlation of the sentiment with other macroeconomic factors like Gas and Oil Price was established.

Academics from the University of Bristol have published two papers with analysis of periodic patterns in daily media content and consumption under the ThinkBIG project [17]. The first paper [6] was focused on the scrutiny of 87 years of the United States and United Kingdom newspapers between 1836 and 1922. Studies have found people's behavior were strongly correlated with the weather and seasons. In the second paper [7], presented at 2016 IEEE International Conference on Data Mining, the authors pay their attention to discovering mental health changes. The team analyzed Twitter content in the United Kingdom and Wikipedia access over four years using data mining and sentiment analysis techniques. They found that negative sentiment tends to be overexpressed in the winter with the peak value in November, while more aggressive emotions like anxiety and anger seem to be overexpressed between September and April. To conclude, both papers states that people's collective behavior follows strong periodic patterns.

This paper describes the development of the program for monitoring peoples' mood through Twitter content in Russian. This paper aims at implementing the software product for exploring the temporal and geographical mood patterns in Russia using machine learning techniques. In contrast with issues mentioned above, this program is designed to process Twitter data in the online mode, i.e. to receive data directly from Twitter API in real time, rather than analyze the pre-collected messages corpus.

The paper is organized as follows. In section 2 the program implementation, methodology, and data collection are described. Section 3 is focused on results and further ways of research. The limitations of this paper are provided in section 4.

2. Implementation, data, and methodology

With the popularization of social media, particularly the micro-blogging website Twitter, a vast amount of content is directly generated by people every day. In addition to textual information, which seems to have affective component, Twitter

messages are also time-stamped and geo-located. Consequently, both tweets meaning and extended information about a message can be analyzed in a purpose of scientific studies in general and in the exploration of public mood variations particularly.

The client-server model was implemented for this project, where the server-side application collects and analyzes Twitter content, and the client-side web application visualizes results. Python was selected as a preferred programming language because of its cross-platform operability, open source code and a vast number of third-party libraries. The Google App Engine [9] cloud platform was used to run and host this project on Google's infrastructure in Python runtime environment. The applications data is stored in Google App Engine Cloud Datastore [4], that is, a high-performance database.

Fig. 1 illustrates the process of public mood monitoring; it's clear that it can be divided into several parts. Firstly, messages obtained via Twitter API [19] using Python-based client library. Secondly, the identification of a federal subject for each obtained message is performed. Thirdly, sentiment analysis is executed. Fourthly, the information of emotional polarity of messages is stored in the database. At the last step, the client-side application visualizes results. Details for these parts are given in the following sections.

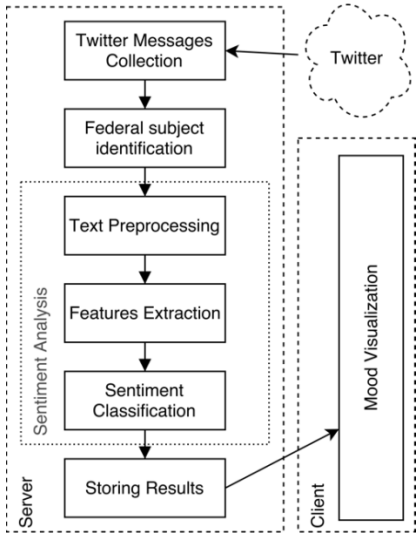

Fig. 1. The program architecture

2.1 Twitter messages collection

Twitting with a location is the geotagging feature in the Twitter platform. On the one hand, this feature helps to provide more meaningful experience for users by making messages more contextual. On the other hand, it makes possible for

researchers to analyze Twitter content from the location-based point of view. In order to use the Tweeting with location feature users must opt-in, i.e. turn location "on". The location will be displayed with users Tweets only in case if they give explicit permission for location extraction. Twitter tracks their location via mobile geo-services or IP.

It's common for IT companies to release its API to the public so that other software developers can design products that are powered by its service. To access Twitter content programmatically it's necessary to register the developer application in Twitter Developers Console. Using credentials from the registered application it's possible to interact with Twitter API from the code of the program. The open-sourced library Tweepy [18] was used in this project to communicate with the Twitter platform and use its' API. The cron job, that is, time-based job scheduler in Unix-like computer OS, is searching and collecting new tweets in Russian with geotagging information via Tweepy every minute. In other words, the information about newly published messages is updated in the program every minute.

2.2 Federal subject identification by message coordinates

For each message collected at the previous step the administrative-territorial entity should be defined according to ISO 3166-2:RU standard, that is, part of ISO 3166 standard published by the International Organization for Standardization, which describes the principal subdivisions of all countries coded in ISO 3166-1.

Due to high implementation complexity, it was decided to use existing geographical services to identify federal subjects' codes. The GeoNames [8] worldwide geographical database was selected for identification of the federal subject code by message latitude and longitude values. This service provides developers with HTTP REST API, which includes identification of the country ISO code and the administrative subdivision of any given point. According to GeoNames terms and conditions of use, there are 30000 requests daily limit and 2000 hourly limit for the code identification functional.

2.3 Sentiment Analysis

The sentiment analysis process can be divided into three steps. At the first step, text preprocessing for collected messages is executed to prepare textual information for sentiment analysis. At the second step, classification features are extracted from prepared messages. At the last step, sentiment classification for each message is performed. The detail description of the steps is as follows.

1) Text preprocessing

Texts generated by humans in social media sites contain lots of noise that can significantly affect the results of the sentiment classification process. Moreover, depending on the features generation approach, every new word seems to add at least one new dimensional, that makes the representation of texts is sparse and high-dimensional, consequently, the task of the classifier has become more complex.

According to [10], text preprocessing has been found crucial on the sentiment classification performance.

To prepare messages, such text preprocessing techniques as reverting repeated letters, removing URLs, removing numbers, converting to lowercase, word normalization and stemming were used in this program. Removing and replacing tasks was performed using regular expressions. The morphological analyzer PyMorphy2 [11]was used for words normalization. Stemming of normalized words was performed using NLTK Python library [16].

2) Features extraction

A basic step for a static natural language processing task tends to be the conversion of raw text into features, which provides a machine learning model with a simpler, more comprehensible view of the text. The bag-of-words model was used to calculate texts embedding using unigrams and bigrams.

3) Sentiment classification

In this project, the multinomial Naïve Bayes classification algorithm for binary sentiment analysis task was used because of its tendency to perform significantly well in the texts classification task and wide usage [20], [2], [14]. The basic idea of Naïve Bayes technique is to find the probabilities of classes assigned to texts by using the joint probabilities of words and classes [5]. Consider the given data point x and class $c \in C$. The starting point is Bayes' theorem for conditional probability which estimates as follows:

$$P(c|x) = \frac{P(x|c)}{P(x)}$$

$$P(x|c) = \frac{count(x,c)}{count(c)}$$

Where *count(x, c)* is the count of word x in class c; *count(c)* is a count of all words in class c. For texts with unknown words, the estimation (2) might be problematic because it would give zero probability. The usage of Laplace smoothing is a common way to solve this problem (3).

$$P(x|c) = \frac{count(x,c)+1}{count(c)+|V|+1}$$

Where $|V|$ is the length of vocabulary in training set.

From the assumption of word independence, it appears that for data point $x = \{x_1, x_2, \ldots, x_i\}$ the probability of each of its features to occur in the given class is independent. Thus, the estimation of this probability can be calculated as follows:

$$P(c|x) = P(c)\prod\prod P(x_i|C)$$

In this context, that means the final equation for the class chosen by a naive Bayes classifier is (5).

$$c_{nb} = \underset{c \in C}{\operatorname{argmax}}\, P(c)\prod\prod P(x_i|c)$$

To avoid underflow and increase speed, the Naive Bayes calculations are performed in the log space (6).

$$c_{nb} = \underset{c \in C}{\operatorname{argmax}} \left(\log \mathrm{P(c)} + \sum \log P(x_i|c) \right)$$

The Naive Bayes classifier was trained on the corpus of short texts in Russian based on Twitter messages [3], which consists of 114991 positive and 111923 negative tweets. The 10-fold cross-validation shows accuracy up to 83%.

2.4 Storing results

Every time the cron job have been executed, the new information about publication time, the amount of positive and negative messages for each federal subject is stored in the database.

2.5 Visualization

To explore temporal public mood variations and location based mood values the website was implemented. Both types of graphics were developed with the framework Google Charts [1], which provides developers with the tool for constructing interactive charts for browsers and mobile devices. There are three graphics displayed at the website. The first one is the Russia mood geo chart, where the current mood state for each federal subject is visualized. The second one and the third one are temporal mood changes plots through the day and through the week respectively.

1) Mood variations

The information about the time of the day and day of the week is extracted from messages to calculate temporal mood changes. Next, the public mood changes are calculated using the following equitation:

$$mood_t = \frac{pos_t}{pos_t + neg_t}$$

Where, pos_t is the number of positive messages in the specific period *t*; neg_t is the number of negative messages in the specific period *t*. The temporal mood changes chart through the day and through the week are plotted in the program. These charts are constructed over all data that have been processed by the program already, so the level of its accuracy and reliability increases with the number of analyzed tweets.

2) Mood geo chart

To plot the mood geo chart, for each federal subject the mood values are calculated using (7) for the last hour. Next, the federal subjects in the geo chart are marked with colors from green to red, where green color means the predominance of positive tweets; yellow color means the balance between the amount of positive and negative messages; red color means the predominance of negative tweets. Fig. 2 illustrates the example of the public mood geo chart for Russia.

Fig. 2. Example of the public mood geo chart for Russia

3. Results

As a result, the program for public mood monitoring through Twitter content in Russian is implemented as web-service, which can be found by the URL http://twittermood-ru.appspot.com/. The program collects new messages, which are published on Twitter, in real time mode, performs sentiment analysis, process the data obtained at the previous step, and visualizes the results. The mood geo chart provides with an opportunity for monitoring mood values in different regions of Russia for the last hour. The other plots offer valuable insights about temporal public mood changes based on all collected data.

The further research will be focused on extending of analyzed feelings, that means, monitoring not only positive or negative sentiment expressions, but also the expression of fear, sadness, joy, and anger. In addition, the multiclass sentiment classification can be implemented to enhance the quality of public mood calculations.

4. Limitations

Despite a wide range of Twitter content analysis benefits, it also has some drawbacks. Technically, Twitter users are not representative of the public, consequently, tweets are not representative of the public opinion [15]. Findings in this article apply only to the population of Twitter users geo-located in the Russia. In this work, it's possible to make claims only about the population of Russia Twitter users and not the general population.

References

[1]. "Charts | Google Developers," *Google Developers*. [Online]. Available: https://developers.google.com/chart/. [Accessed: 18-Mar-2017].

[2]. R. Collins, D. May, N. Weinthal, and R. Wicentowski, "SWAT-CMW: Classification of Twitter Emotional Polarity using a Multiple-Classifier Decision Schema and Enhanced

Emotion Tagging," *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pp. 669–672, 2015. – 2015.
[3]. "Corpus of short texts in Russian," *Julia Rubtsova*. [Online]. Available: http://study.mokoron.com/. [Accessed: 18-Mar-2017].
[4]. "Datastore - NoSQL Schemaless Database | Google Cloud Platform," *Google Cloud Platform*. [Online]. Available: https://cloud.google.com/datastore/. [Accessed: 18-Mar-2017].
[5]. L. Dey, S. Chakraborty, A. Biswas, B. Bose, and S. Tiwari, "Sentiment Analysis of Review Datasets Using Naïve Bayes' and K-NN Classifier," *International Journal of Information Engineering and Electronic Business*, vol. 8, no. 4, pp. 54–62, Aug. 2016.
[6]. F. Dzogang, T. Lansdall-Welfare, and N. Cristianini, "Discovering Periodic Patterns in Historical News," *Plos One*, vol. 11, no. 11, 2016.
[7]. F. Dzogang, T. Lansdall-Welfare, and N. Cristianini, "Seasonal Fluctuations in Collective Mood Revealed by Wikipedia Searches and Twitter Posts," *2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW)*, 2016.
[8]. "GeoNames," *GeoNames*. [Online]. Available: http://www.geonames.org/. [Accessed: 18-Mar-2017].
[9]. "Google App Engine Documentation | App Engine Documentation | Google Cloud Platform," *Google Cloud Platform*. [Online]. Available: https://cloud.google.com/appengine/docs/. [Accessed: 18-Mar-2017].
[10]. E. Haddi, "Sentiment analysis: text, pre-processing, reader views and cross domains," dissertation, 2015.
[11]. M. Korobov, "Morphological Analyzer and Generator for Russian and Ukrainian Languages," *Communications in Computer and Information Science Analysis of Images, Social Networks and Texts*, pp. 320–332, 2015.
[12]. S. Kumar, S. Maskara, N. Chandak, and S. Goswami, "Empirical Study of Relationship between Twitter Mood and Stock Market from an Indian Context," *International Journal of Applied Information Systems*, vol. 8, no. 7, pp. 33–37, 2015.
[13]. T. Lansdall-Welfare, F. Dzogang, and N. Cristianini, "Change-Point Analysis of the Public Mood in UK Twitter during the Brexit Referendum," *2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW)*, 2016.
[14]. B. Le and H. Nguyen, "Twitter Sentiment Analysis Using Machine Learning Techniques," *Advanced Computational Methods for Knowledge Engineering Advances in Intelligent Systems and Computing*, pp. 279–289, 2015.
[15]. A. Mitchell and P. Hitlin, "Twitter reaction to events often at odds with overall public opinion," *Pew Research Center*, vol. 4, 2013.
[16]. "Natural Language Toolkit," *Natural Language Toolkit — NLTK 3.0 documentation*. [Online]. Available: http://www.nltk.org/. [Accessed: 18-Mar-2017].
[17]. "thinkBIG – Patterns in Big Data: Methods, Applications and Implications," *thinkBIG*. [Online]. Available: http://thinkbig.enm.bris.ac.uk/. [Accessed: 18-Mar-2017].
[18]. "Tweepy," *Tweepy*. [Online]. Available: http://www.tweepy.org/. [Accessed: 18-Mar-2017].
[19]. "Twitter Developer Documentation — Twitter Developers," *Twitter*. [Online]. Available: https://dev.twitter.com/docs. [Accessed: 18-Mar-2017].
[20]. Y. Wan and Q. Gao, "An Ensemble Sentiment Classification System of Twitter Data for Airline Services Analysis," *2015 IEEE International Conference on Data Mining Workshop (ICDMW)*, 2015.

Программа для мониторинга общественных настроений в России на основе сообщений из Twitter

С.И. Сметанин <sismetanin@gmail.com>
Национальный исследовательский университет «Высшая школа экономики», 101000, Россия, г. Москва, ул. Мясницкая, д. 20

Аннотация. Ежедневно пользователями социальных сетей генерируются значительные объемы текстового контента, который дополнительно содержит информацию о координатах и времени публикации. Эти данные могут быть проанализированы и использованы для оценки общего состояния большой популяции пользователей с целью решения научных вопросов из широкого спектра дисциплин. В данной статье описывается разработка программы для мониторинга общественных настроений на основе анализа тональности сообщений из русскоязычного сегмента социальной сети Twitter с использованием методов машинного обучения. В разработанном программном продукте была использована многоуровневая сетевая архитектура «клиент-сервер». Написанное на Python серверное приложение собирает сообщения пользователей через Twitter API, осуществляет предварительную обработку текста, анализирует эмоциональную окраску сообщений с использованием мультиномиального наивного Байесовского классификатора и определяет их принадлежность к административно-территориальным субъектам страны. Клиентское веб-приложение визуализирует результаты анализа тональности, которые состоят из карты настроений России, где для каждого административно-территориального субъекта указывается текущий показатель настроения, а также из графиков изменения настроения в течение дня и в течение недели. В процессе разработки программного средства были задействованы облачные сервисы. Серверная часть была развернута на платформе Google App Engine, которая позволяет выполнять веб-приложения на серверах Google, то есть полностью абстрагироваться от инфраструктуры, поэтому при работе сервер не нуждается в администрировании. Данные программы хранятся в облачной базе данных Google Cloud Datastore, которая полностью интегрирована с Google App Engine.

Ключевые слова: анализ тональности; общественные настроения; социальные сети

DOI: 10.15514/ISPRAS-2017-29(4)-22

Для цитирования: Сметанин С.И. Программа для мониторинга общественных настроений в России на основе сообщений из Twitter. *Труды ИСП РАН*, том 29, вып. 4, 2017 г., стр. 315-324 (на английском языке). DOI: 10.15514/ISPRAS-2017-29(4)-22

Список литературы

[1]. Charts | Google Developers. *Google Developers* (online). Доступно по ссылке: https://developers.google.com/chart/. [Дата обращения: 18.03.2017]

[2]. R. Collins, D. May, N. Weinthal, and R. Wicentowski, “SWAT-CMW: Classification of Twitter Emotional Polarity using a Multiple-Classifier Decision Schema and Enhanced Emotion Tagging,” *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pp. 669–672, 2015. – 2015.

[3]. Corpus of short texts in Russian. *Julia Rubtsova*. (online). Доступно по ссылке: http://study.mokoron.com/. [Дата обращения: 18.03.2017]
[4]. Datastore - NoSQL Schemaless Database | Google Cloud Platform. *Google Cloud Platform* (online). Доступно по ссылке: https://cloud.google.com/datastore/. [Дата обращения: 18.03.2017]
[5]. L. Dey, S. Chakraborty, A. Biswas, B. Bose, and S. Tiwari, "Sentiment Analysis of Review Datasets Using Naïve Bayes' and K-NN Classifier," *International Journal of Information Engineering and Electronic Business*, vol. 8, no. 4, pp. 54–62, Aug. 2016.
[6]. F. Dzogang, T. Lansdall-Welfare, and N. Cristianini, "Discovering Periodic Patterns in Historical News," *Plos One*, vol. 11, no. 11, 2016.
[7]. F. Dzogang, T. Lansdall-Welfare, and N. Cristianini, "Seasonal Fluctuations in Collective Mood Revealed by Wikipedia Searches and Twitter Posts," *2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW)*, 2016.
[8]. GeoNames. *GeoNames* (online). Доступно по ссылке: http://www.geonames.org/.
[9]. Google App Engine Documentation | App Engine Documentation | Google Cloud Platform. *Google Cloud Platform* (online). Доступно по ссылке: https://cloud.google.com/appengine/docs/. [Дата обращения: 18.03.2017]
[10]. E. Haddi, "Sentiment analysis: text, pre-processing, reader views and cross domains," dissertation, 2015.
[11]. M. Korobov, "Morphological Analyzer and Generator for Russian and Ukrainian Languages," *Communications in Computer and Information Science Analysis of Images, Social Networks and Texts*, pp. 320–332, 2015.
[12]. S. Kumar, S. Maskara, N. Chandak, and S. Goswami, "Empirical Study of Relationship between Twitter Mood and Stock Market from an Indian Context," *International Journal of Applied Information Systems*, vol. 8, no. 7, pp. 33–37, 2015.
[13]. T. Lansdall-Welfare, F. Dzogang, and N. Cristianini, "Change-Point Analysis of the Public Mood in UK Twitter during the Brexit Referendum," *2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW)*, 2016.
[14]. B. Le and H. Nguyen, "Twitter Sentiment Analysis Using Machine Learning Techniques," *Advanced Computational Methods for Knowledge Engineering Advances in Intelligent Systems and Computing*, pp. 279–289, 2015.
[15]. A. Mitchell and P. Hitlin, "Twitter reaction to events often at odds with overall public opinion," *Pew Research Center*, vol. 4, 2013.
[16]. Natural Language Toolkit. *Natural Language Toolkit — NLTK 3.0 documentation* (online). Доступно по ссылке: http://www.nltk.org/. [Дата обращения: 18.03.2017]
[17]. thinkBIG – Patterns in Big Data: Methods, Applications and Implications. *thinkBIG*. (online). Доступно по ссылке: http://thinkbig.enm.bris.ac.uk/. [Дата обращения: 18.03.2017]
[18]. Tweepy. *Tweepy* (online). Доступно по ссылке: http://www.tweepy.org/. [Дата обращения: 18.03.2017]
[19]. Twitter Developer Documentation — Twitter Developers. *Twitter*. (online). Доступно по ссылке: https://dev.twitter.com/docs. [Дата обращения: 18.03.2017]
[20]. Y. Wan and Q. Gao, "An Ensemble Sentiment Classification System of Twitter Data for Airline Services Analysis," *2015 IEEE International Conference on Data Mining Workshop (ICDMW)*, 2015.