Usage of Sentiment Analysis to Tracking Public Opinion

Zoia Kochuieva, Natalia Borysova, Karina Melnyk and Dina Huliieva

National Technical University "Kharkiv Polytechnic Institute", Kirpichova, 2, Kharkiv, 61002, Ukraine

Abstract

This study reveals the problems of analysis of public opinion. The description, use cases and efficiency estimation of software for sentiment analysis of public opinion have been presented. The relevance of the problem of sentiment analysis as one of the important tasks of computational linguistics is substantiated. An overview of the existing classical methods of sentiment analysis and some software applications that solve this problem is conducted. The business process model of analysis of public opinion is presented in the form of BPMN-diagram. The principles of operation of the developed classifier that used the lexicon-based method are described. The model of determining the tonality of the news in the form of an activity diagram was considered. The efficiency estimation of the developed lexicon-based classifier has been evaluated based on standard metrics (Recall, Precision). The obtained results have been compared with values of similar metrics based on the using of the Naïve Bayesian Classifier and Recurrent Neural Network Cmeans Classifier. The calculation of the Recall and Precision has been conducted for two cases: the sentiment analyzer used a dictionary of affective words without slang words and with slang words. Conducted numerical studies show increasing of the efficiency of the sentiment analyzer by 5-6% in the case of using a dictionary with slang words.

Keywords[1]

Sentiment analysis, sentiment analysis methods, lexicon-based sentiment analysis, sentiment analysis software, automated analysis of public opinion, classifier efficiency estimation

1. Introduction

The problem of public opinion analysis today falls within the interests of many professionals, including marketers, sociologists, political scientists and many others. Public opinion is a form of mass consciousness, which reflects the attitude (hidden or overt) of different groups of people to the events and processes of society that affect their interests and needs. Public opinion has expressed publicly. It affects the functioning of society and political system. At the same time, public opinion is a set of many individual opinions on a specific issue that concerns a group of people. The structure of public opinion includes mass moods, emotions, feelings, as well as evaluations and judgments. In addition, public opinion is a base for a government for the following: an idea of the interests of the population, attitudes to innovations, events, statements of officials, politicians, public figures, mechanisms for presenting the most acute and significant problems for citizens, and others. People at present can express their opinions on the Internet, and the number of statement grows every day. The manual analysis of the opinions is not possible, because the public opinion can change quickly. So, there is an urgent need to automate the process of public opinion analysis. Opinion mining is a research domain dealing with automatic methods of detection and extraction of opinions and sentiments presented in text [1]. This study focuses on sentiment analysis, which can determine the emotional attitude of the author of the statement to any entity (a product, service, the person, the organization, an event) and / or its properties, signs, parts, etc.

COLINS-2021: 5th International Conference on Computational Linguistics and Intelligent Systems, April 22–23, 2021, Kharkiv, Ukraine
EMAIL: aliseiko@gmail.com (Z. Kochuieva); borysova.n.v@gmail.com (N. Borysova); karina.v.melnyk@gmail.com (K. Melnyk); dgulieva@ukr.net (D. Huliieva)
ORCID: 0000-0002-4300-3370 (Z. Kochuieva); 0000-0002-8834-2536 (N. Borysova); 0000-0001-9642-5414 (K. Melnyk); 0000-0001-8310-745X (D. Huliieva)

2. An overview of existing methods and tools of sentiment analysis

Consider the classic methods and some software applications for sentiment analysis that currently exist.

2.1. A synopsis of methods of sentiment analysis

All methods of automated sentiment analysis can be divided into the following groups:

1. Rule-based methods.
2. Lexicon-based methods.
3. Supervised machine learning methods.
4. Unsupervised machine learning methods.
5. Hybrid methods.

Rule-based methods use sets of rules identified by experts based on the analysis of texts in the subject area. The information system (IS) defines the tone of the texts based on these rules. To obtain high accuracy of the classifier, it is necessary to write a large number of rules. Nevertheless, it is a long and time-consuming process. In addition, the rules describe only specific domain, so the changing the domain needs the re-composing of the rules. However, this approach is most accurate with a good rule base, because rule-based algorithms are closely related to word semantics. Also, these methods give good results in the classification of structured or poorly structured texts, such as texts of scientific articles, or other grammatically correct texts without spelling errors. However, rule-based methods depend heavily on the language of the texts, i.e. they are not universal [2].

Lexicon-based methods use affective lexicons to analyze texts. A tonal dictionary is a list of words with tonality for each one (positive, negative, neutral) and weight coefficients (for example, from -5 to 5, or from -10 to 10). The IS analyzes some text, finds particular words from the dictionary, calculate the overall tone of the whole text according to the weights of these words. There are many methods of calculation the tone of the text, for instance, the using of arithmetic mean. However, these methods are not universal, because they depend on the language of the texts, as well as on the domain area (each domain area nedds own dictionary) [3].

Supervised machine learning methods for training the classifier use a training sample (texts corpora). This set contains from marked texts divided into classes. The classifier or IS can determine the tonality of new texts unknown based on this sample. The most widely used methods of sentiment analysis are the naive Bayesian classifier and the algorithm of support vector machine. The usage of supervised machine learning methods gets good results; the accuracy of the algorithms can exceed 90%. The main difficulty of using these methods is creating a test sample to teach the classifier, because the quality of texts corpora has an influence on the effectiveness of the classifier [4].

Unsupervised machine learning methods for training the algorithm use a training sample (corpora) based on undivided into classes and unmarked texts. The biggest weights allow find the most common words in the text, however, they are presented only in a limited number of texts of the whole set. One of the most used method in practice is the K-means algorithm. However, this group of methods for determination the tonality of the texts is not frequently used because of lower accuracy in comparison with supervised machine learning methods [4].

Hybrid methods are a combination of methods of different groups. They allow using advantages of the selected methods and eliminating their disadvantages. An example of such hybridization is the method of sentiment analysis, which accommodates the syntactic structure of the text and the relationship between words in a sentence. The classifier applies such text structures that used to express a person's emotional attitude toward an object. The decision tree and the lexicon-based method utilize simultaneously for it. It should also pointed out that the dictionaries can consist of positive, negative words and inverter words. Inverter words are the words that can change the polarity of the whole sentence. The nodes of the tree are the words of the sentence. The values of the higher node are calculated on the following: the values of the lower nodes, the ability of the word to invert the tonality, and the tonality of the word from the dictionary. If IS ignores the sentence structure, it can get the wrong classification result. For example, the attitude to the news can be defined as negative because of two negative words and one positive, while the attitude is neutral based on the content of the message [5].

2.2. Existing software for sentiment analysis

In addition to the existing methods of sentiment-analysis, some sentiment-analysis software has analyzed in the work. This software is based on different approaches for solving the problem and is designed for using in different conditions. Each software has a number of advantages and disadvantages. In order to define best software solution for sentiment analysis, independent organizations and experts create reviews with lists of TOP-10 or TOP-12 sentiment-analysis tools based on surveys of a large number of users. Sometimes such lists are differ, but some tools are presented in all reviews. So, analysis of existing software contains from such tools.

As the developers say, Rosette Sentiment Analyzer has a machine learning model that was training on tweets and reviews to detect strong positive and negative sentiments in documents. It also uses an entity extraction to identify set of products from customer review, where customer mentioned two or more products. Rosette has sentiment analysis and entity extraction models for six languages. However, user can add new languages for training Rosette. Rosette Text Analytics is the company owner of Rosette Sentiment analyzer. It has several price plans for customers: Analytics, Full Stack, Enterprise. All these plans include sentiment analysis feature. There are three subplans within the Analytics Plan: Starter for $100 per month, Medium for $400 per month and Large for $1,000 per month. The Full Stack Plan has two subplans: Small for $500 per month and Medium for $1,350 per month. The pricing for Enterprise Plan is revealed upon request [6].

Social Searcher is a free social media search engine. It could be used by users in two possible ways: firstly, for searching in social networks (such as Twitter, Facebook, Youtube, Instagram, Flickr, Vimeo, etc) in a real time, and secondly, for the monitoring of social media. Social Searcher gives such information about posts: sentiment, type of content and language. Its syntax supports phrase searching and operators using. Social media monitoring could be made with Social Searcher API. API concentrates information about brand mentions and provides access to it. This information could be sort by date or popularity, could be filter by social network, sentiment or content type, could be found from chosen posts, could be export to CSV format, etc. Users' data is stored till their subscription is valid. There are two types of users that could work with Social Searcher: Free and Premium [7, 8]. Social Searcher could be used for free with 100 searches during the day and 2 email alerts. It has three price plans: Basic for 3,49 € per month with 200 searches during the day, 3 email alerts, 3 monitorings, 3000 posts per month, all mentions in the web; Standard for 8,49 € per month with 400 searches during the day, 5 email alerts, 5 monitorings, 20000 posts per month, all mentions in the web; Professional for 19,49 € per month with 800 searches during the day, 10 email alerts, 10 monitorings, 100000 posts per month, all mentions in the web. And now there is a special offer on their site "Start Standard plan 14-day free trial" [9].

Repustate's sentiment analysis multilingual API uses a combination of machine learning methods to identify sentimental insights in messages from all possible communication channels and users' data. There are five steps of natural language processing for sentiment analysis by Repustate:

Step 1: POS-tagging.

Step 2: Lemmatization.

Step 3: Prior polarity determining and intensity of the polarity calculating.

Step 4: Determining of negations, amplifiers and other grammatical constructs.

Step 5: Machine learning using.

Repustate offers two price plans: Standard for $299 per month, that provide English language processing only, document sentiment only, standard document volume, basic support by email, cloud API; Custom is available upon request and provide all 23 supported languages processing, document, topic and aspect sentiment analysis, expanded document volume, premium support by phone and email, cloud API / on-premise deployment, customized machine learned models, named entity recognition, data retrieval (news, social, blogs), sentiment analysis dashboard, video/audio/image content retrieval, enterprise semantic search [10].

Following the information from website, Social Mention is a special social media platform for searching and collecting users' content from the web. Social Mention monitors more than 100 social networks properties. It provides searching, analysis and daily alerts in social media, third-party APIs and applications. Developers can interact with the Social Mention website using special API [11. 12].

Social Mention gives the results by four characteristics: strength, sentiment, passion, reach. Strength is the likelihood of mentioning a certain brand in social networks during the last 24 hours. Sentiment is the ratio of all positive mentions to all negative mentions. Passion is a likelihood of multiple mention of brand by same people. Reach is a measure of the influence diapason. It is a ratio of number of brand mention by unique authors to the total number of mentions [13]. Users can work with API for free if they make less the 100 requests during a day. Usage of Social Mention for commercial purposes is required of contacting with the developers [12].

MeaningCloud's Sentiment Analysis API is a tool for making a detailed attribute-leveled aspect-based multilingual sentiment analysis of different texts. It separates texts into three classes: positive, negative and neutral texts. Aspect-based analysis means that polarity value for the whole text calculated according to polarity values of all sentences of this text and relationships between them. API could be useful for facts and opinions extraction, irony identification, polarity disagreement finding, etc. It is possible to work with API using users' sentiment dictionaries and users' sentiment models [14]. Customers can use MeaningCloud's Sentiment Analysis API for free and analyze 20000 requests per month with free support and SaaS deployment. There are also four paid plans exist: Start-Up Plan for $99 monthly with 120000 requests per month, standard support and SaaS deployment; Professional Plan for $399 monthly with 700000 requests per month; Business Plan for $999 monthly with 4200000 requests per month; Enterprise Plan for custom paid per month with custom requests per month, premium support, SaaS and On-premises deployment [15].

In addition, we do not overlook the sentiment-analysis software of such global IT giants as IBM, Microsoft and Google.

IBM Watson Natural Language Understanding (NLU) allows detecting the insights in structured and unstructured data. The NLU simplifies the text analysis for metadata extracting from content, which includes concepts, keywords, categories, entities, semantic roles and relations. The NLU is a good application to recognize emotions and sentiments, because it returns emotion and sentiment for the whole text and keywords in the text for deeper analysis. The IBM Watson NLU uses Watson Knowledge Studio to understand the texts in nine languages. The NLU also has the conversation feature that enables to build and deploy chatbots and virtual agents across a different communication channels. It provides the infrastructure for matching with individual use cases, therefore it gives users the support they need [16]. The page [17] demonstrates the necessary information about the price and even link for pricing calculator. It is worth noting that IBM Watson can be also used for free.

Microsoft Azure Cognitive Service Text Analytics API supplies advanced processing of unstructured natural language texts. The API has four main features: Sentiment Analysis (and Opinion Mining), Key Phrase Extraction, Language Detection and Named Entity Recognition. The API uses classification methods for Sentiment Analysis. Sentiment score is a numeric score between 0 and 1. If the score value close to 1, text is positive. If the score value close to 0, text is negative. English, French, Spanish and Portuguese languages are supported and 11 additional languages in preview. The API uses techniques from Microsoft Office's sophisticated Natural Language Processing toolkit for Key phrase extraction. English, German, Spanish, and Japanese languages are supported. Key phrases are used for topic detection. The API can detect the language of text for 120 languages. The language detection score is a score between 0 and 1. If the score value close to 1, language is detected 100% certainty [18]. Text Analytics can be purchased in tiers [19]. Free Plan allows doing 5000 transactions free per month with three of four main features without Named Entity Recognition. Standard Plan has the same features as a Free Plan, but the quantity of analyzed text records bigger and price for their processing depends on the quantity. S0-S4 plans have all four main features including Named Entity Recognition and cost from $ 74,71 per month to $ 4999,99 per month.

Google Cloud Natural Language API uncovers the structure and text meaning by using machine learning models in a REST API. It could be used for finding mentions about people, places, events, etc., in texts and documents. It allows understanding sentiment about brand or/and product on social media or analyzing customer conversations holding in a call center or a messengers. It searches useful insights on product approbation or user experience from customer conversations in email, chat or social media. It filters inappropriate content and classifies documents by topics; builds relationship graphs of entities extracted from news or Wikipedia articles and extracts tokens and sentences and then identifies parts of speech to create dependency parse trees for each sentence. The Google Cloud Natural Language API supports 11 languages [20]. The API usage is based on the following principle: pay only for the features

you use [21]. Free Plan allows using free all features for 5000 units. If the text contains less than 1,000 Unicode characters, it could be considered as one "unit". Prices in other plans depend on the units' quantity and features, features differ in price, the more units the cheaper.

The analysis has showed that all considered software are multifunctional, but only two of them support the Ukrainian language. Other products allow downloading own model for sentiment analysis and / or dictionary of sentiment words, but this service is paid or developers have set restrictions on the use of models and user dictionaries. Thus, the development of its own sentiment-analyzer of public opinion for Ukrainian-language texts is an urgent task.

3. The model of the sentiment-analyzer of public opinion

In this study, it is proposed to conduct the process of determining the tonality of the news using the lexicon-based method. The main idea of this method is to use the tonal dictionaries, where each word has a certain weight coefficient or several weight coefficients. The calculation of the overall tonality of the whole text is based on the weight coefficients of the words from the dictionary. The dictionary of words tonality has been made for developed sentiment analyzer. Calculations of the tonality of the text have been carried out according to the methodology proposed in [22]. The research [22] demonstrates the determining the tonality of the news for English-language texts. However, the authors pointed out the possibility of using their methodology for other languages based on an appropriate tonality dictionary. Thus, consider the use of the proposed methodology for Ukrainian-language texts.

Let's assume N is a set of news, which is needed for determination the tone according to the comments on them. Denote s_i^N as the tonality of i-th news, $i \in N$.

Let's denote W as a set of words and collocations of the tonality dictionary, so w_j $(w_j \in W)$ – j-th word of this dictionary. Each word has its own tonality, so denote s_j^W as the tonality of w_j-th word from the dictionary W. The range of changes of tonality is measured in the range $[-100; 100]$, where negative values characterize the negative tonality, and positive values are positive tonality, respectively. If the word w_j occurs in the text of the comment with a negation, then it is necessary to use formula (1). The efficiency of formula (1) is proved in [22]:

$$s_j^{W\prime} = \begin{cases} \max\left(\dfrac{s_j^W + 100}{2}, 10\right), & s_j^W < 0 \\ \min\left(\dfrac{s_j^W - 100}{2}, -10\right), & s_j^W \geq 0 \end{cases}. \quad (1)$$

Denote I as set of words-intensifier, for example: дуже, трохи, доволі, etc. Some words have a positive intensification, then they belong to the subset $I_P \subset I$, and words with negative intensificationare contained in the subset $I_N \subset I$, respectively.

Let's denote K as the set of comments to all news from the set N, then K_i is the subset of comments to the i-th news, $i \in N, K_i \subset K$. Denote $s_{ki}^C (i \in N, K_i \subset K)$ as the tonality of the k-th comment of the i-th news. Different groups of people who are public speakers can write comments. There are three categories of comments: the opinion of the media (the opinions of authors of articles in various online publications about particular news); the opinion of the people (the opinions of ordinary citizens about news); the opinion of experts (the opinions of people, who are the experts in domain related with given news). Let's suggest $K_i^c \subset K_i$ as a subset of the comments of the c-th category:

$$\bigcup_{c=\overline{1,3}} K_i^c = K_i.$$

Then $s_i^{N_c}$ is the tonality of the i-th news in the c-th category. In this paper, it is proposed to determine $s_i^{N_c}$ and s_i^N by formula (2) and (3), respectively:

$$s_i^{N_c} = \frac{1}{r_c} \sum_{k,i} s_{ki}^C, k \in K_i^c, i \in N, \quad (2)$$

$$s_i^N = \frac{1}{3}\sum_{c=\overline{1,3}} s_i^{N_c}, \quad (3)$$

where r_c is the cardinality of the set K_i^c.

To determine the tone of the comment s_{ki}^C, it is necessary to find $W_k, (W_k \subset W)$, where W_k is the set of words of the comment k-th. The words from W_k are the elements of the set W and the sets of words-intensifier of the comment k-th I_P^k $(I_P^k \subset I_P)$ and $I_N^k (I_N^k \subset I_N)$ simultaneously, if they exist for the k-th comment. The cardinality of the sets $\left|I_P^k\right| = q_{Pk}$ and $\left|I_N^k\right| = q_{Nk}$, respectively. If all selected words have only one tonality, for example, positive, then the whole comment is considered like positive one. In doing so, some methods for determining the tonality offer to find A_P – the arithmetic mean of all positive words from k-th comment by formula (4) or A_N – the arithmetic mean of negative words by formula (5):

$$A_P = \frac{1}{p}\sum_j s_j^W, j \in W_k, \quad (4)$$

$$A_N = \frac{1}{n}\sum_j s_j^W, j \in W_k, \quad (5)$$

where p and n are the number of positive and negative words in the k-th comment, respectively. Thus, the tonality of the comment s_{ki}^C is defined as follows:

$$s_{ki}^C = \begin{cases} A_P, & \forall\, j \geq 0, j \in W_k \\ A_N, & \forall\, j < 0, j \in W_k' \end{cases} \quad (6)$$

The paper [22] empirically shows the inaccuracy of estimating the tonality of a sentence or text by arithmetic mean. Authors of this methodic propose its own version of determining the tonality of the comment sentence.

Consider a model for determining the tonality of news based on the calculation of the tonality of the set of comments to this news, using the model from [22].

Let's consider additional variables: X_P and X_N are the overall positive and negative sentiment in k-th comment respectively; E_P and E_N are the overall positive and negative evidence in k-th comment respectively.

$$X_P = \min\left\{\frac{A_P}{2 - \lg(3.5p + q_{Pk})}, 100\right\}, \quad (7)$$

$$X_N = \max\left\{\frac{A_N}{2 - \lg(3.5n + q_{Nk})}, -100\right\}, \quad (8)$$

$$E_P = \min\left\{\frac{A_P}{2 - \lg(3.5p)}, 1\right\}, \quad (9)$$

$$E_N = \max\left\{\frac{A_N}{2 - \lg(3.5n)}, -1\right\}. \quad (10)$$

These variables are needed to determine the tonality of a particular comment (Fig. 1). The process of estimation of tonality of news is shown in the form of the activity diagram using the activity element "Defining the tone of the k-th comment", while the parameters are X_P, X_N, E_P and E_N. Thus, the model of determining the tonality of the news in the form of the diagram has considered.

Let's consider the process of tracking public opinion based on using of sentiment analysis in more details. To develop effective classifier for a specific domain, it is necessary to create a model of this process. There are many techniques and case tools for modelling business process. This research propose to use Business Process Modeling Notation (BPMN) for formalizing the process of tracking public opinion. The Fig. 2 presents the business process model of the given process in the form of BPMN-diagram. To start working with the sentiment analyzer, the user has to add a news item. Then administrator or other user has to add comments with defined category for this news.

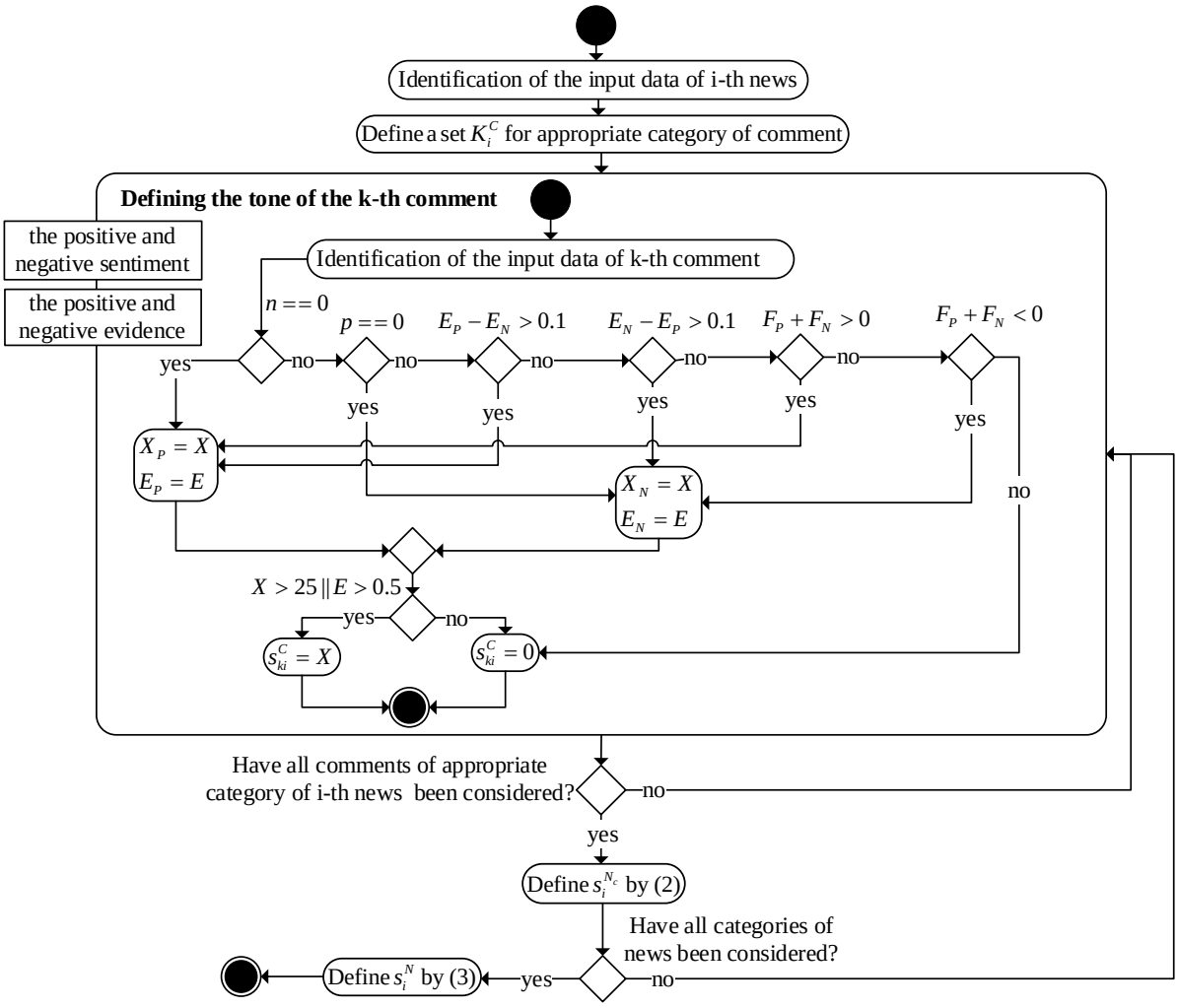

Figure 1: The model for determining the tone of the news

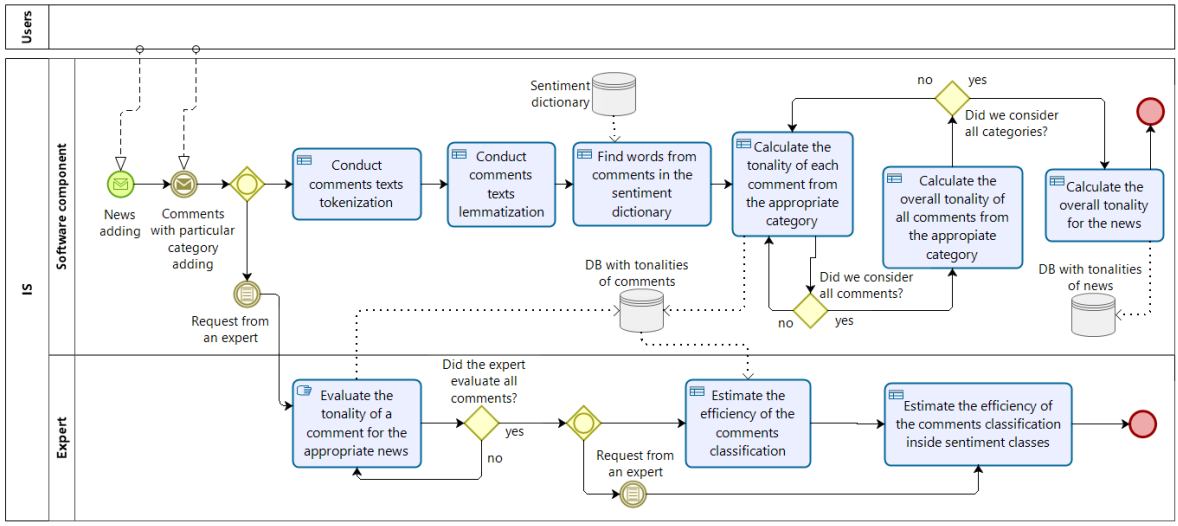

Figure 2: The BPMN-diagram of the given business process

Next step is the tokenization and lemmatization for each comment. This stage of analysis allow comparing the found words with the words available in the sentiment-dictionary. If the word is in the dictionary, its weight coefficient is taken for calculations. The tonality of each comment further is

calculated according to the proposed model of determining the tonality of the news. Next step leads for calculation of the tonality of the comments with particular category. The purpose of the step is determination of the attitude of different public opinion leaders to each news item. Finally, classifier estimates the overall tonality for each news item. It means the general tonality of public opinion about the news. To assess the efficiency of the work of the sentiment analyzer according to the standard metrics Recall and Precision, it is necessary to ask the tonality of all comments from experts of considered domain area. Detailed information of this process is presented in paragraph 4 of this article.

Let's consider the functional and non-functional requirements for the sentiment analyzer. There are three roles of user: administrator, user and expert. The administrator can add and delete news, comments, comment categories and user accounts, as well as view the tonality of comments and news and the results evaluating the effectiveness of the sentiment analyzer. The user has the ability to add news and comments to them, view all the sentiment assessment results and the effectiveness assessment results. The expert has the ability to manually set own assessment of the sentiment of comments and view the results of assessing the effectiveness of the program.

Non-functional requirements include the following: intuitive and user-friendly interface, reliability of data transfer and storage, usability, performance, high performance. The whole functionalities of the developed sentiment analyzer for different categories of users are presented in the form of a use-case diagram in Fig. 3.

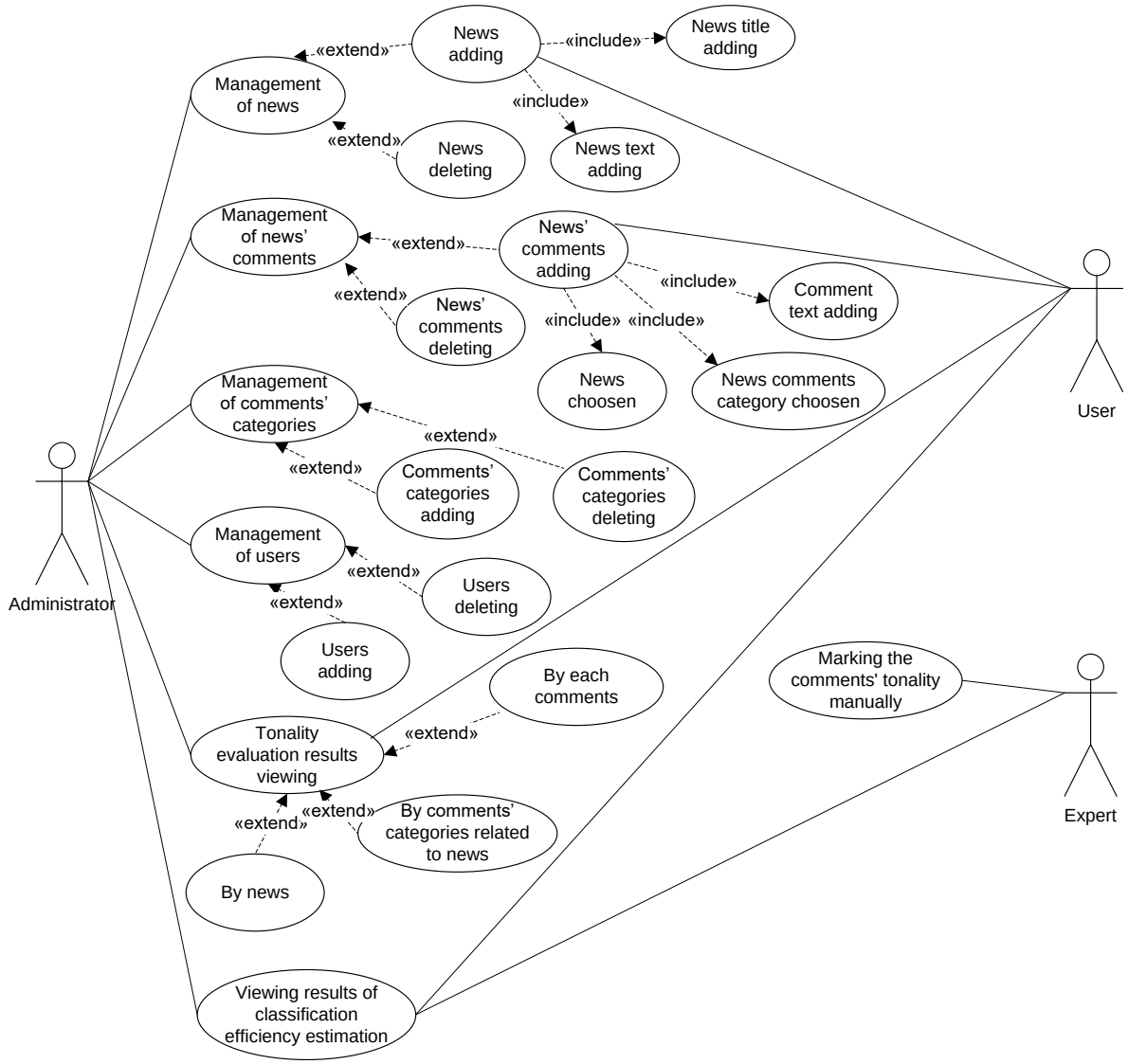

Figure 3: The functional possibilities of the sentiment analyzer

All data and results of the sentiment analyzer work are stored in the database. The logical structure of the database allow seeing the relationships between different objects or entities of domain area. Every business rule of the work of the sentiment analyzer should be base for creating of model of the database. There are many different models of a database of domains, although the Entity-Relationship Model is the most widely used one. Let's consider the database model for the sentiment analyzer (Fig. 4).

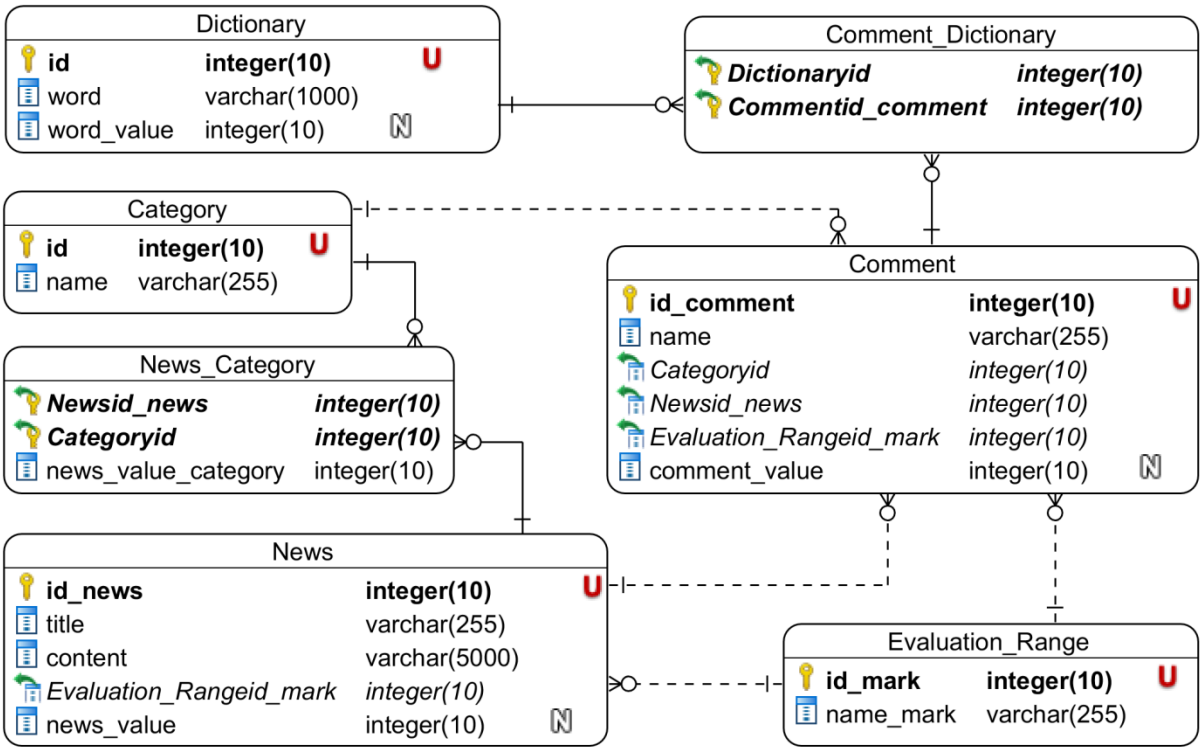

Figure 4: The model of the database

Database model consists of following entities:

– the entity «Dictionary» represents list of sentiment words with their weight coefficients;

– the entity «News» describes all news added by all users with their tonality expressed both in words and numerical value;

– the entity «Comment» represents all comments for all news with their categories and tonality expressed both in words and numerical value;

– the entity «Category» describes all comments' categories;

– the entity «Evaluation_Range» represents all tonality expressions by words;

– the associative entity «Comment_Dictionary» describes the set of words from the comment presented in a dictionary;

– the associative entity «News_Category» describes tonality evaluation expressed by numerical value for each comments' categories that belong all news.

Description of these entities and their attributes is presented in the Table 1.

Table 1

Description of database model

Entity	Attribute	Attribute description
Dictionary	id	Word's id
	word	Word
	word_value	Word's weight coefficient
News	id_news	News' id
	title	News' title
	content	News' content
	Evaluation_Rangeid_mark	News' general tonality expressed by words

	news_value	News' general tonality expressed by numerical value
Comment	id_comment	Comment's id
	name	Comment's content
	Categoryid	Comment's category
	Newsid_news	Comment related news id
	Evaluation_Rangeid_mark	Comment's tonality expressed by words
	comment_value	Comment's tonality expressed by numerical value
Category	id	Category id
	name	Category name
Evaluation_Range	id_mark	Tonality evaluation id
	name_mark	Tonality evaluation expressed by words
Comment_Dictionary	Dictionaryid	Word's id
	Commentid_comment	Comment's id
News_Category	Newsid_News	News' id
	Categoryid	Category id
	news_value_category	The numerical value of tonality evaluation

4. The efficiency estimation of the developed analyzer

According to the aforementioned information, the standard metrics Precision and Recall have been used to evaluate the efficiency of the sentiment analyzer work. To calculate these metrics, it is necessary to find the following indicators:

- true positive – the number of answers we expected to see and received at the exit;
- false positive – the number of answers that we did not expect to see, but the analyzer mistakenly returned them at the exit;
- false negative – the number of answers that we expected to see, but the analyzer did not return them at the exit;
- true negative – the number of answers that we did not expect to see, and the analyzer did not return them at the exit.

The Table 2 presents the examples of the assessments of the sentiment analyzer and experts of several comments with different categories related to one news item, and the matching between results.

Table 2

The results of assessments of tonality of comments made by expert and sentiment-analyzer

№	Evaluation of the program	Expert assessment	Match status
1	negative	negative	+
2	positive	positive	+
3	positive	negative	-
4	negative	negative	+
5	negative	negative	+
6	negative	negative	+
7	positive	very positive	-
8	positive	positive	+
9	negative	negative	+
10	positive	positive	+
11	negative	positive	-
12	negative	negative	+
13	positive	positive	+

14	very positive	very positive	+
15	positive	neutral	-
16	negative	negative	+
17	very negative	very negative	+
18	very negative	very negative	+
19	positive	positive	+
20	positive	positive	+
21	very positive	positive	-
22	positive	positive	+
23	positive	positive	+
24	positive	positive	+
25	positive	positive	+
26	positive	negative	-
27	very positive	very positive	+
28	positive	positive	+
29	positive	positive	+
30	positive	positive	+

Precision is calculated as the proportion of relevant responses in the total volume of all responses issued by the sentiment analyzer by the formula:

$$Precision = \frac{TP}{TP + FP} \tag{11}$$

Recall is calculated as the proportion of relevant responses in the total number of relevant responses. Recall is calculated by the formula:

$$Recall = \frac{TP}{TP + FN} \tag{12}$$

The Table 3 presents examples of evaluation by the sentiment analyzer of the same several texts of comments that are presented in the Table 2, but the results of the evaluation are presented in the relevant classes. The one point in the Table 2 indicates class of the text obtained by the sentiment analyzer. If there is a letter in parentheses next to the one point, it means that the sentiment analyzer made a mistake and incorrectly assigned the text to this class. The letter in parentheses indicates the correct class for this text chosen by the expert.

Table 3

The results of distribution the texts on classes by the sentiment analyzer

№	Text evaluation				
	Very negative (nn)	Negative (n)	Neutral (N)	Positive (p)	Very positive (pp)
1		1			
2				1	
3				1(n)	
4		1			
5		1			
6		1			
7				1(pp)	
8				1	
9		1			
10				1	
11		1(p)			
12		1			
13				1	
14					1

15			1(N)		
16		1			
17	1				
18	1				
19			1		
20			1		
21				1(p)	
22			1		
23			1		
24			1		
25			1		
26			1(n)		
27				1	
28			1		
29			1		
30			1		

According to the results of calculations, the following metric values have been obtained by (11)-(12): Precision = 0.861; Recall = 0.849. Moreover, the value of the Precision metric within the class has range from 0.807 to 0.921, and the Recall metric – from 0.775 to 0.930. Such results indicates that the sentiment analyzer works adequately in general and within each tonality class as well.

The obtained values of metrics Precision and Recall of the lexicon-based classifier have been compared with the values of such metrics for two other classifiers: Naïve Bayesian Classifier and RNN Cmeans Classifier, based on Recurrent Neural Network. Results for Naïve Bayesian Classifier and RNN Cmeans Classifier are taken from [23]. All results of classifiers efficiency evaluation are presented in the Table 4.

Table 4

Efficiency estimation of classification results

Metrics	Lexicon-based Classifier	Naïve Bayesian Classifier	RNN Cmeans Classifier
Precision	0,861	0,869	0,878
Recall	0,849	0,853	0,870

The authors of the research [23] describe an experiment where additional training of classifiers on Slang corpus, which the tonality of slang words have been marked up, allowed to increase the efficiency of classifiers by 10-11%. Therefore, based on this information, we decided to supplement our dictionary with slang words from the dictionary [24] and re-examine the work of the developed lexicon-based sentiment analyzer. The results of metric calculations for the same three classifiers are presented in the Table 5. Results for Naïve Bayesian Classifier and RNN Cmeans Classifier are also taken from [23].

Table 5

Efficiency estimation of classification results with Slang dictionary and Slang corpus

Metrics	Lexicon-based Classifier	Naïve Bayesian Classifier	RNN Cmeans Classifier
Precision	0,915	0,975	0,982
Recall	0,895	0,948	0,965

Comparing the results of calculations of metrics from the Tables 3 and 4, the increasing of the efficiency of the work of the lexicon-based classifier has not happen by 10-11%, the increasing has occurred by 5-6%. This can be explained by the fact that not every of the analyzed comments contain slang words, they are used only in the comments, that the opinion of the people reflect.

5. Conclusions

The paper presents an approach to solving the problem of evaluation of public opinion using the sentiment analyzer. The existing methods and software for sentiment analysis are analyzed. The comparative characteristics of these methods have allowed choosing the lexicon-based methods. The proprietary algorithm for solving the problem of determining the attitude of public opinion representatives based on their comments to news is proposed. To calculate the sentiment of the comments, the technique was used, which was first applied to Ukrainian-language texts, and its own dictionary of sentiment words, subsequently supplemented with slang words. A functional model of the business process of tonality identification by sentiment analyzer and a database model as well as their description are presented. All its functionality has shown in the form of a use-case diagram and described. The efficiency of the developed sentiment analyzer was assessed using the standard Precision and Recall metrics. A comparative analysis of the efficiency of the Lexicon-based Classifier and Naïve Bayesian Classifier, RNN Cmeans Classifier has been carried out. It is shown that adding the of slang words to the sentiment dictionary increases the efficiency of the Lexicon-based Classifier by 5-6%, while additional training of two other classifiers on Slang corpus showed an increase in efficiency by 10-11%.

6. References

[1] S. Ion, C. Bucur, Applying Supervised Opinion Mining Techniques on Online User Reviews. Informatica Economica Journal. 16. URL: https://core.ac.uk/download/pdf/27056535.pdf

[2] D. Vilares, C. Gómez-Rodríguez, M. A. Alonso, Universal, unsupervised (rule-based), uncovered sentiment analysis, Knowledge-Based Systems, volume 118, 2017, pp. 45–55, URL: https://www.sciencedirect.com/science/article/pii/S0950705116304701. doi: 10.1016/j.knosys.2016.11.014

[3] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, M. Stede, Lexicon-based methods for sentiment analysis. Computational Linguistics. 37. (2011) 267–307. doi: 10.1162/COLI_a_00049

[4] Z. Rahimi, S. Noferesti, M. Shamsfard, Applying data mining and machine learning techniques for sentiment shifter identification. Language Resources and Evaluation, volume 53, issue 2, 2019, pp. 279–302. doi: 10.1007/s10579-018-9432-0

[5] K. Ravi, V. Ravi, A survey on opinion mining and sentiment analysis: Tasks, approaches and applications, Knowledge-Based Systems, volume 89, 2015, pp. 14–46, URL: https://www.sciencedirect.com/science/article/pii/S0950705115002336. doi: 10.1016/j.knosys.2015.06.015

[6] Rosette Sentiment Analyzer. URL: https://www.rosette.com/capability/sentiment-analyzer/#overview

[7] About Social Searcher. URL: https://www.social-searcher.com/about/

[8] Social Searcher API V2.0 Released. URL: https://www.social-searcher.com/2015/08/04/social-searcher-api-v2-0-released/

[9] Social Searcher pricing. URL: https://www.social-searcher.com/pricing/

[10] Sentiment analysis. Unlock the meaning in your data. URL: https://www.repustate.com/sentiment-analysis/

[11] About Social Mention. URL: http://socialmention.com/about/

[12] Social Mention API. URL: http://socialmention.com/api/

[13] Social Mention. Frequently Asked Questions. URL: http://socialmention.com/faq

[14] MeaningCloud's Sentiment Analysis API. URL: https://www.meaningcloud.com/developer/sentiment-analysis

[15] MeaningCloud pricing. URL: https://www.meaningcloud.com/products/pricing

[16] IBM Watson Natural Language Understanding. URL: https://www.predictiveanalyticstoday.com/ibm-watson-alchemyapi/

[17] How NLU pricing works. URL: https://www.ibm.com/cloud/watson-natural-language-understanding/pricing

[18] Microsoft Azure Cognitive Service Text Analytics API. URL: https://www.predictiveanalyticstoday.com/microsoft-azure-text-analytics-api/
[19] Cognitive Services pricing – Text Analytics API. URL: https://azure.microsoft.com/en-us/pricing/details/cognitive-services/text-analytics/
[20] Google Cloud Natural Language API. URL: https://www.predictiveanalyticstoday.com/google-cloud-natural-language-api/
[21] Google Cloud. Cloud Natural Language. URL: https://cloud.google.com/natural-language/ pricing
[22] A. Jurek, M. D. Mulvenna, Y. Bi, Improved lexicon-based sentiment analysis for social media analytics. Security Informatics. 4, 9 (2015). URL: https://security-informatics.springeropen.com /articles/10.1186/s13388-015-0024-x#article-info doi: 10.1186/s13388-015-0024-x
[23] N.V. Borysova, K.V. Melnyk, Efficiency estimation of methods for sentiment analysis of social network messages, Bulletin of National Technical University "KhPI", Series: System Analysis Control and Information Technologies. 2 (2019) 76–81. doi:10.20998/2079-0023.2019.02.13
[24] N. V. Borysova, V. V. Niftilin, Avtomatyzovane stvorennia elektronnogo slovnyka, in: E. I. Sokol (Eds.), Proceedings of XXV International scientific-practical conference in Information technologies: science, engineering, technology, education, health, MicroCAD-2017: Part 1 (May 17–19, 2017), NTU "KhPI", Kharkiv, 2017. p. 32