I.J. Modern Education and Computer Science, **2022, 1, 1-16**
Published Online February 2022 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijmecs.2022.01.01

The Technique of Key Text Characteristics Analysis for Mass Media Text Nature Assessment

Oksana Babich
The Taras Shevchenko National University, Kyiv, 03680, Ukraine
E-mail: o.babichknu@gmail.com

Viktor Vyshnyvskiy
The State University of Telecommunications, Kyiv, 03110, Ukraine
E-mail:vish_vv@ukr.net

Vadym Mukhin
National Technical University of Ukraine Igor Sikorsky Kyiv Polytechnic Institute, Kyiv, 03056, Ukraine
E-mail: v_mukhin@i.ua

Irina Zamaruyeva
The State University of Telecommunications,
Kyiv, 03110, Ukraine

Michail Sheleg
The State University of Telecommunications, Kyiv, 03110, Ukraine
E-mail: shelegmike@gmail.com

Yaroslav Kornaga
National Technical University of Ukraine Igor Sikorsky Kyiv Polytechnic Institute, 03056, Ukraine

Received: 07 July 2021; Revised: 02 September 2021; Accepted: 31 October 2021; Published: 08 February 2022

Abstract: The paper presents the technique for analysis of text emotional nature which is a key characteristic of Mass media news text. Emotions inherent design its Emotional coloring and become a significant feature of mass media news texts. The technique proposed measures the degree of exposure of emotions and allocates them by rating. Emotional coloring is defined by emotional characteristics and by grammar categories, and a set of rules is applied to regulate wordforms interaction. Techniques for verbal units analysis are examined. The Heavy Natural Language Processing models and Machine learning techniques are considered. They are compared and the optimum one is defined to resolve the problem of Emotional coloring evaluation. A system prototype is developed on the basis of this technique. It allocates news by influence rating according to their key parameters. The examples of texts' emotional nature recognition results by means of the prototype are presented. The visualization of emotional nature analysis results highlights additional features of the news text's emotional nature and expresses them in numeric values. It is exposed both by sentences and by the whole news text, with tracking of news Emotional coloring dynamics. The results presented have application in analysis procedure intending to studying Mass media, particularly informational environment with concomitant factors, and their impact on political and social interrelation.

Index Terms: Analysis, text nature, procedure, emotional coloring, assessment, machine learning, technique.

1. Introduction

Analysis of Mass media news content is an important part of information processing to study a social environment state and moods inherent. Emotions detection (ED) is a vital component for Mass media text nature assessment and its main characteristics definition. Emotions that are intrinsic to mass media news text impact both their news content and public sentiments of their audience correspondingly. All this entails related behavioral reactions with the appropriate

outcomes for a social environment. This essential factor stipulates the importance of Mass media research for a social environment study. The results of research presented in this paper provide a multilateral assessment of Mass media news content, with a focus on the intensity of emotions manifestation in the text, the Emotional coloring, and an overall Intensity of information impact of a text. The components listed herein take into account key text characteristics of Mass media news that can be evaluated and measured with our technique. They form the text nature that is a significant factor that can influence human perception and behavior as a consequence.

Mass media research opens the vast field of notions that needed to be put into formal view for natural and social procedures and to be described by formal language. This way of description is a tool to make possible machine reading of the specific information, with the view of identification of semantics of natural language texts, particularly Mass media news.

The article presents the technique for ED and subsequent Mass media news texts and information workflow evaluation. The related content is presented as follows: Section 2 highlights some basic concepts of text-based Emotions detection, with their related approaches and models. Section 3 presents developed procedures for key text characteristics formal description and ED phases with analysis of text nature. Techniques for ED, analysis of their characteristics and their comparative indexes for key system functions, and models for ED in a text are stated in Section 4. And Section 5 concludes the article.

The problem of text nature assessment within data analysis for machine learning operations is being examined in different aspects of social life. Among them, the research of the target audience as an important factor was examined and such characteristics as types of audiences on the internet, and their reactions to social networks content, a relation of different audiences' activeness to publications, and these factors presentation in the graphical form were outlined in [1]. Analysis of verbal expressions types inherent to the Internet, and means of their marking for recognition are studied in [2]. Techniques for multi-domain sentiment analysis and learning concept polarity are developed in [3]. The aspects of data processing by different Sentiment analysis methods and their results classification with a distribution of their polarities by ratings are stated in [4]. Improvement for features classification with their optimization to evaluate emotional attitude is developed in [5].

The results of the researches abovementioned reveal diverse aspects of text nature evaluation including assessment of audience parameters and text nature evaluation by mean of sentiment polarities distribution. But evaluation of emotional characteristics of texts is missed in these approaches. So, the following research and development results in Emotions detection fill up the gap.

The important issue is that today mass media are based on computer platform, specially on the distributed computer systems [6 – 11].

2. Emotions Detection Basic Concepts. Text-based Approaches and Models Related

Emotions detection approaches highlight the rule construction, Machine learning (ML), and hybrid approaches as the general approaches to detecting emotions from texts. Significant drawbacks and strengths are associated with each approach.

For example, evaluation of deriving emotional responses of individuals in the process of their interaction with a certain environment is examined in [12], with analysis of social media content (in particular, tweet) characteristics and inherent emotion and their ratio. Three approaches to comparison of sentiment analysis undertaken to collate the sentiment and emotion present in tweet text are described there. Results of various statistical tests are visualized to assess the significance of any differences in the assignment of each positive/negative/neutral tweet by each of the 3 presented therein methods. The techniques were correlated with visualization of intermethod reliability of tweet assignment into each category.

The rule-based approach outlines major grammatical and logical rules to follow in order to detect emotions from documents. The rule construction approach encompasses keyword recognition (KR) and lexical affinity methods. The KR method deals with the construction or the use of emotion dictionaries or lexicons. The task is to find occurrences of these search words in a written text at the sentence level. Once the keyword is identified within the sentence, a label is assigned to the sentence [13].

The review of recent studies [13] reveals the main approaches adopted by researchers in the design of text-based ED systems. Some recent state-of-the-art proposals are discussed further in the article. Their approaches employed, results obtained, and their strengths and weaknesses are stated. The direction of further research in this field and essential opportunities for improving the detection of emotions from texts are also mentioned.

The state of being emotional is often aligned with making conscious arousal of feelings subjectively or with influence from the environment, thus emotions such as happiness, sadness, fear, anger, surprise, and so on are derived from the personal (subjective) experiences of individuals and as well as their interactions with their surroundings.

A brief survey outlined hereinafter demonstrates the current techniques employed for the detection of emotions from text [13]. As a rule, a dataset which is their vital component is formed with verbal structures and a set of emotions, based predominantly on Paul Ekman's classification and some additional principles. Major publicly available datasets

and their features are discussed below [13]. Their databases consist both of wordforms and sentences extracted from texts of versatile nature. The related researches results are the following [13].

The International Survey on Emotion Antecedents and Reactions (ISEAR) database constructed by the Swiss National Centre of Competence in Research and lead by Wallbott and Scherer consists of seven emotion labels (joy, sadness, fear, anger, guilt, disgust, and shame).

The Semantic Evaluations (SemEval) is a database formed of headlines extracted from news websites. The data in this database are rich in emotional content for emotion extraction and it is labeled using the 6 emotional categories (ie, joy, sadness, fear, surprise, anger, and disgust) presented by Paul Ekman. Its dataset consists of over 10 000 sentences annotated dimensionally in accordance with the Valence-Arousal-Dominance (VAD) emotion representation model. These sentences were obtained from news headlines, essays, blogs, newspapers, fiction, letters, and travel guides of writers and readers, thus spanning a wider domain. A subset of the dataset has also been annotated categorically using Ekman's basic emotion model making it suitable for dual representational designs.

Related systems are functional for social media analysis and their datasets and other components are characterized by the following [13].

The Workshop on Computational Approaches to Subjectivity, Sentiment, and Social Media Analysis (WASSA-2017) was constructed to detect emotion intensities in tweets. Its data was annotated for four discrete emotions, including joy, anger, fear, and sadness.

Smile dataset. This dataset contains discrete emotion annotations of anger, disgust, happiness, surprise, and sadness. Data for emotion analysis can also be obtained from other social media Application Programming Interfaces (APIs) such as Facebook, Google, Twitter, and so on, and also from blog posts.

Cecilia Ovesdotter Alm's Affect data. The dataset was constructed from tales and annotated categorically for fear, angry, disgust, sad, happiness and surprise. It also contains other annotations as feeler, intensity, and lists of emotion words characterized for assisting emotion annotations.

Daily Dialog. This dataset was built by crawling dialogues from regular human conversations. Its contains sentences annotated for neutral, anger, disgust, fear, happiness, sadness, surprise discrete emotion labels.

Emotion-Stimulus data. The data are built from sentences containing emotions as well as factors that cause emotions (emotion stimulus). The emotion-labeled sentences are annotated in compliance to the Ekman's basic emotion categories plus shame.

Crowdsourcing dataset. It is annotated for 13 emotions, that is, surprise, happiness, sadness, anger, fun, worry, love, hate, enthusiasm, boredom, relief, emptiness, and neutral.

As for current state-of-the-art text-based proposals, they had been juxtaposed with emotions on the Russell's Circumplex Model of Affect in a knowledge extraction process in order to determine the tagged texts carrying emotions. The emotion-carrying texts were then checked for negations and features selected according to their frequency of occurrences. After a series of procedures all texts assigned sentiments were classified into one of four emotion classes as “happy-active,” “happy-inactive,” “unhappy-active,” and “unhappy-inactive.” They had also classified a large quantum of tweets into a restricted number of four classes; this implied that most tweets will not be classified into their specific emotion categories but their nearest categories as per the classes available which may result in inaccurate classifications. Furthermore, emojis that have been found to convey rich emotion content were ignored as their model focused only on text classification [13].

As for machine learning techniques employed for ED, the following new detection methods are outlined hereinafter [13]. A similarity technique based on a vector similarity measure (VSM) and STASIS method is presented by Mozafari and Tahayori. They then compared the performances of their methods with the Keyword method and concluded that their proposed method outperformed the Keyword method under standard conditions. Their work initially implements the Stasis for extracting semantic relationships from texts and then determines the similarity measure of texts using the VSM method.

Detection of cyber abuse and emotion of aggression particularly was performed by Aditya Malte and Pratik Ratadiya. Using a bidirectional transformer-based BERT Architecture, they detected this from Facebook multilingual texts, English, Hindi, and a mixture of both languages (Hinglish) texts [13]. During the pre-training phase, the model was trained for masked language model (Masked ML) and next sentence prediction tasks. After fine-tuning, the data were categorically classified into three, that is, covertly aggressive (CAG), overtly aggressive (OAG), and non-aggressive (NAG).

Matla and Badugu explored the performance of K-nearest neighbors (K-NN) algorithm and Naïve Bayes Machine learning (NB ML) techniques in the detection of emotions using tweets in the Sentiment corpus. Their approach showed that under constant conditions, the NB outperformed the KNN with an accuracy of 72.06% as compared to 55.50% [13].

The effects of emoji inclusion in detecting emotions from texts were explored by LeCompte and Chen. Their work adopted the Keyword Identification method and the Support vectors machine (SVM) and Multinomial Naïve Bayes (MNB) to classify emotions into sad, angry, scared, happy, surprised, thankful, and love. Their result indicated an improvement in performance where emojis were considered rather than where they were not considered. They also indicated that under equal conditions, MNB performed better than SVM [13].

A hybrid framework for the automatic detection of multilanguage text data was presented by Jian et al. Their proposed framework used Natural language processing (NLP) techniques to extract emotions existing in texts, and then classified them per the concept of emotion models presented by Ekman using the SVM first and then NB. As they assert their method provided better results in comparison with other methods achieving an accuracy of 72.81%; however, a robust ML Classifier can be used for better performance [13].

Nida et al [13] developed an automatic emotion classifier for tweets using three emotion corpora where features are initially extracted from these databases individually and synonyms of emotion words were generated with the WordNet dictionary in their databases. The model was trained using the SVM classifier and emotions were further classified into one of the six Ekman's categories of emotion. The related work showed an accuracy of 59.71%, 63.24%, and 67.86% on the three emotion corpora used, respectively.

The Emotex model [14] is designed for detecting emotions from texts using a supervised learning method and emotion dictionaries. Their approach consists of two methods: an offline and an online classification task. The example of the Emotex system for creating models for classifying emotions. Emotex was built using emotion-labeled texts from Twitter and the SVM, NB, and decision tree classifiers. The data went through preprocessing and feature vector constructions to obtain the training datasets for the classification model. The online approach used the model created in the offline approach to classify live streams of tweets in real-time. This model is simple and describes a wide range of emotional states, with 28 affect words [14]. It is juxtaposed to the Circumplex model and as a result, emotions revealed become distributed in a two-dimensional circular space, containing pleasure and activation dimensions. Four major classes of emotions are considered: Happy Active, Happy-Inactive, Unhappy-Active, and Unhappy Inactive. So, it is relevant to Twitter messages to reveal and assess properly its audience moods.

The survey identifies that the use of text-based ED had not been adequately explored for the purpose of analytics with predictive bias. As we can see from the review of models for ED, the predominant part of them is capable to reveal a limited spectrum of emotions, mainly 5-13. The models are capable to classify detected emotions into 2-4 degrees of intensity and distribute emotions into major classes that demonstrate psychological state, from Active to Inactive.

All approaches abovementioned lack thorough evaluation of emotional characteristics of texts. Particularly, evaluation of Mass media news texts characteristics that form public sentiments and concomitant outcomes. The means specified are unable to detect a vast spectrum of emotions that are studied in modern psychological research and are recognized as a vital characteristic of a human being. Emotions can be a trigger for certain behavioral manifestations and are important to be detected. Besides, the models listed above don’t reveal such parameters as humor, and malicious joy, and irony as its derivative manifestation.

As a rule, the data for analytical purposes are designated for specific missions performing. Besides, such purposes have situational nature and variable demands for information and the data related. As a consequence, the amount of data for the specific task may be insufficient and needs to be filled up. That’s why it is critical to use an adequate technique for data processing within the ED procedures. Features inherent to the set of emotions in this research are also to be taken into consideration to be processed in a proper way.

With the view of Mass media news exploration that corresponds to our goals, a new processing model is necessary to be used for the purpose of Emotions detection. A procedure for ED and a thorough evaluation of emotional characteristics of Mass media news texts must provide rapid processing of data stored in a defined format and containing a multilateral information. And a rapid addition of new data to the system is obligatory due to a variable political and social environment, that impacts the content to be processed.

For this purpose, it is important to define key characteristics that are subject to be examined. Analysis aspects of the research above-stated lack all-around analysis of emotions nature in verbal structures and their properties for the purpose above-stated. With this in view, the research theses presented herein state the means of formal description of Mass media news texts study, particularly their emotional characteristics recognition and its general intensity assessment.

The results obtained in the process of text-based ED research are presented herein.

The methods examined are used for processing Mass media news texts features with the view of their formal description.

The main text characteristics to be the subject of formal description are:

- the components that form the Emotional coloring of the news text;
- the factors that condition the power of the text information impact.

3. Key Text Characteristics Formal Description. ED Procedure Phases for Text Nature Analysis

The procedure for analysis of text nature which is transmitted by emotions nominating wordforms includes the following.

Evaluation of Text Emotional coloring (EC) which encloses three main phases. They are:

Phase 1.

The intensity of emotions is the essential characteristic, that is revealed in a degree of exposure. Each of the emotions interested is analyzed by experts in the scope of 16 points, and allocated by its intensity in the range from –8 to +8 that provides a vast scope of emotional colors to be evaluated. The Semantic differential technique is proposed for the distribution of each emotion intensity by the range. A list of emotions is evaluated by experts according to their scales-factors set.

The procedure composes the following:

- the collection of expert judgments for the list of emotions;
- working out of expert's questionnaires, calculation of the average for each emotion from the list and as a result – the average for each emotion from all experts' assessment.

The values obtained are the basis for attribution of own weighting ratio to each emotion for the subsequent evaluation of the news text, particularly its emotional coloring, based on own research survey and psychology workout study [15]. The results are allocated in Table 1, where emotions are marked with corresponding rates according to their degree of exposure, arranged by alphanumerical code, and allocated by their basic characteristics. These parameters are the key to define text features, particularly their emotional nature.

Table 1. Allocation of Emotions intensity characteristics

№	Alpha code of emotion	Emotion	Degree of exposure
1.	AQ	Calming	1
2.	AS	Sadness	–1
3.	AD	Distrust	–2
4.	T	Calming down	2
5.	I	Confidence	2
6.	AV	Regret	–3
7.	AJ	Irritation	–3
8.	S	Envy	–4
9.	R	Satisfaction	4
10.	V	Malevolence	–5
11.	U	Astonishment	5
12.	AT	Approval	5
13.	AK	Disappointment	–5
14.	AU	Anxiety	–6
15.	ZA	Hope	6
16.	AF	Hatred	–7
17.	AI	Joy	7
18.	AR	Fear	–8
19.	O	Euphoria	8

The above-stated arrangement serves for ED in the process of the identification of the emotions by wordforms in the news text, which is the core of Phase 2.

Input data for Phase-2 encloses the alphanumerical code which is a part of the basic rules.

The basic rules include:

- logical-semantic rules with regulations for emotions wordforms interaction;
- a set of procedures for text processing and emotional coloring and text nature assessment.

To regulate emotional identification by wordforms that transmit them, the next classification (Table 2) is proposed. It includes both the above-mentioned and numerical codes of sentence syntactic and semantic categories and rules.

Table 2. Numerical codes of sentence semantic categories and lexical-semantic rules

№	Code	Sentence semantic category
1.	60*	Object of emotion
2.	61*	Matter of emotion
3.	62*	Subject of emotion
4.	63*	Cause of emotion
5.	64*	Source of information
6.	70*	Influence on emotion modulus
7.	71*	Signs of emotion interaction
8.	72*	Merger of signs of emotions
9.	73*	Summarizing
10.	74*	Factor of information impact intensity

Further regulations (Table 3) comprise wordforms interaction inside of the sentence and resulting output [16]. Their use defines the Emotional coloring of the text correctly and more precisely.

Table 3. Logical semantic rules

№	Code	Rules	Explanation
1.	70*	The rule of influence on emotion modulus. Influences intensity change of Emotional coloring of a wordform, to which it interacts.	$k(S_i) = \left\|k(w_j)\right\| + p(w_i),\ p = 0{,}5$ *(1)* where S_i is a syntactic structure, k – is a weighting factor of the Emotional coloring of a wordform; w_j and w_i are lexical units with Emotional coloring which are an integral part of S_i ; p – is a weighting factor that influences wordform modulus.
2.	71*	The rule of signs interaction. When wordforms with different signs of weighting factors (positive or negative) are summed up the sign of wordform with a bigger modulus is the prevalent for a newly formed syntactic structure.	$k(S_i) = k_i$, if $\left\|k(w_i)\right\| > \left\|k(w_j)\right\|$ *(2)* where S_i is a syntactic structure, k – is a weighting factor of Emotional coloring of a syntactic structure; w_j and w_i are lexical units with Emotional coloring which are an integral part of S_i , provided $k(w_j)$ and $k(w_i)$ are located on different intervals.
3.	72*	The rule of merger. Is valid for interacting wordforms with identical signs of weighting factor. A weighting factor of a wordform which has modulus of a bigger value gets unifying for both of them	$k(S_i) = \left\|k_j\right\|$, *if* $k(w_j) \geq k(w_i)$ *(3)* where S_i is a syntactic structure, k – is a weighting factor of Emotional coloring of a syntactic structure; k_j – is a weighting factor of Emotional coloring of a syntactic structure S_i ; w_j and w_i are lexical units with Emotional coloring which are an integral part of S_i , provided $k(w_j)$ and $k(w_i)$ are located on the same interval.

4.	73*	The summarizing rule. Is a logical result of the rules above-stated. Estimates Emotional coloring of each sentence of a text.	$k(R) = \frac{\sum_{i=1}^{N}(S_i)}{N}$ *(4)* where R is an arbitrary sentence of a text; $k(R)$ – is an Emotional coloring weighting factor of an arbitrary sentence of a text; $\sum_{i=1}^{N} k(S_i)$ – is an amount of syntactic structures weighting factors in this particular sentence; N – is a quantity of syntactic structures within a sentence. If $\sum_{i=1}^{N} k(S_i)$ is equal to zero the above-mentioned amount is not divided into N and Emotional coloring of a sentence is neutral.
5.	74*	The rule of information influence intensity assessment. Demonstrates the ratio of a number of values of components that form information impact to their number.	$I = \frac{\sum_{i=1}^{m} G}{n}$ *(5)* where I – is the total intensity of a news information impact which incorporates all the data estimated; G – is the information influence intensity factors which include quantitative and qualitative characteristics of components that form the news nature (a source, a subject and other elements); n – is the number of elements the value degree of which was estimated earlier. This function demonstrates a correlation of an amount of all defined characteristics of news to their number, and shows a news impact degree.

So, these principles are the tool that makes it possible to assess an emotional characteristic of text nature at the lexical level of text organization.

The logical semantic rules above-stated are operated with the set of principles. Syntactic rules which regulate word classes are their integral part.

The Types of clauses that are inherent to modified phrases is the essential component to take into account when using these rules.

The basic and elementary kinds of clauses [17] are canonical clauses and the various kinds of a non-canonical clauses. Among their properties are: a canonical clause is positive, a non-canonical clause is negative. Canonical clause presents the information in the grammatically most basic way. For example, *I have now read most of them* is canonical but *Most of them I have now read* is non-canonical. Besides, non-canonical clause types are expressed with Imperatives (e.g. Please stand up) and exclamations (What a fool I've been!). They are inherent to the Rule of influence on emotion modulus (Rule 70*) when regulating the degree of a wordform.

Word classes or parts of speech (in traditional terms) identification helps us to identify the way of forming the emotional coloring of wordform correctly. It is due to the main of them:

verb: *to be ill, to help*;
noun: *face mask, aid*;
adjective: *disruptive, optimistic*;
adverb: *extremely, clearly*;
coordinator: *or (Hurry or we'll be late), but* (*the task is complicated but it is worth being done);*
subordinator: *if (I wonder if it's true, whether* (*Ask whether their intention is peaceful)*;
interjections: *Wow!, Alas!*

All these word classes are helpful for text nature determination. They contain an emotional pulse and transmit it towards an audience that conduces to a better perception of content.

Some wordforms that require the presence of an auxiliary verb, the most frequent of which involve negation. The construction where *not* is used to negate the verb likewise requires the verb to be an auxiliary:

The reimbursement for damage will be allowed;
The reimbursement for damage will not be allowed.

If there is no auxiliary in the positive, do must be inserted for the auxiliary to form the negative:

This law violated human rights;
This law did not violate human rights.

The negation use also needs auxiliary verbs that belong to both auxiliary and lexical verb classes.

- the non-modal like *be*, *have*, *do*;
- modal verbs like *need, dare*.

All word classes above-stated contribute to form various types of sentences that are subject to recognition. They are helpful to construct a syntactic structure and to define a phrase meaning more precisely. Breaking down a particular text into a syntactic structure and using Logical semantic rules for its adequate apprehension facilitates recognition of its true emotional nature.

4. The Techniques for ED and Their Characteristics Analysis

The techniques used herein for ED and analysis of their characteristics in a text include:

4.1. The models for detection of the article field

This is the first step to split data into two basic categories like is worthy of further processing and unworthy of further processing. The main of them are related to the most popular like politics, business, environment, sports, and entertainment. Though each particular user`s demand for analytics may intake diverse spheres, depending on the case. Training sample and test sample divided selected data in 21760 and 5440, correspondingly.

To envision the system prototype characteristics, like high-speed data processing and memory consumption capabilities, the classical ML methods were chosen and the following models were proved.

Logistics regression
Support vector machine
Support vector machine with stochastic gradient descent
Random Forest
Gradient Tree Boosting

This effort resulted in the decision to use models tf-idf for text transformation and both SVM and SVM with SGD because these models exactly expose the maximum speed on inference (prediction stage). The Logistic regression method results are also close to them employed to accomplish the task assigned. The results obtained are exposed in Table 4.

Table 4. Comparative indexes of text processing models for subject identification

The model	precision	recall	f1-support
SVM	**0.770885**	0.769786	0.769674
SVM with SGD (SGD Classifier)	0.775022	**0.776072**	0.774605
Logistic Regression	0.768380	0.768306	0.767852
Random Forest	0.712008	0.711028	0.719747
Gradient Tree Boosting Machine	0.728642	0.726701	0.726978

Our task requires the following SGD method operational behavior. With the SVM objective function in place and the process of SGD defined, we may now put the two together to perform classification.

SGD operates by using one randomly selected observation from the dataset at a time (different from regular Gradient Descent, which uses the entire dataset together). This is often preferred because it reduces the chances of the algorithm getting stuck in local minima instead of finding the global minimum. The stochastic nature of the points at which the function is evaluated allows the algorithm to "jump" out of any local minima that may exist.

With a time-decaying learning rate (the right-hand plot), the steps get shorter as the algorithm continues to iterate. As the steps get smaller and smaller, the algorithm converges closer and closer to the true value. Centering and scaling data beforehand is another important step for convergence [18 – 20]. So the true value for lexical units is determined more correctly.

4.2. Models for emotions detection in the text

Our research resulted in a decision for emotions identification in the text that includes 2 stages, which are:

- Filtering and finding of words that contain emotions;
- Transformation of words into embedding (vector space) and search of the closest of them by cosine similarity with using of ElasticSearch technique.

This approach makes it possible to process both parts and correct any of them when necessary, without serious impact on another part.

Text processing comprises its layout into sentences. Some categories are sieved, such as stop-words, named entity recognition (NER), wordforms in the upper case, ones without dependent words, and those are not nouns or verbs. Wordforms filtered compose phrases that are to be used in the next stage.

For the second stage of data processing, the techniques of Fast Text embedding and ElasticSearch were used. The FastText embedding technique is preferable here due to its rapid functioning and simultaneous fixation of words both known and unknown. It divides the bulk of words into known parts that solves the out-of-dictionary problem. The approach used is based on the skip-gram model, where each word is represented as a bag of character n-grams. A vector representation is associated with each character n-gram; words being represented as the sum of these representations. This method is fast and allows us to train models on large corpora quickly, allowing us to compute word representations for words that did not appear in the training data [21].

Text Embedding differs from traditional information retrieval techniques. In traditional information retrieval, a common way to represent text as a numeric vector is to assign one dimension for each word in the vocabulary. The vector for a piece of text is then based on the number of times each term in the vocabulary appears. This way of representing text is often referred to as a "bag of words," because we simply count word occurrences without regard to sentence structure. The FastText embedding technique differs from traditional vector representations in some important ways:

- the encoded vectors are dense and relatively low-dimensional, often ranging from 100 to 1,000 dimensions. In contrast, a bag of word vectors is sparse and can comprise 50,000+ dimensions. Embedding algorithms encode the text into a lower-dimensional space as part of modeling its semantic meaning. Ideally, synonymous words and phrases end up with a similar representation in the new vector space;
- sentence embedding can take the order of words into account when determining the vector representation. For example, the phrase "tune in" may be mapped as a very different vector than "in tune".

In practice, sentence embedding often doesn't generalize well to large sections of text. They are not commonly used to represent text longer than a short paragraph [22].

The wordforms processed are transformed into vector space, with the FastText embedding technique. Then the nearest Top-5 vectors in the multidimensional space are defined, functional vector search by cosine similarity in the ElasticSearch is used, resulting in values from (–1) to 1. The technique close to probability calibration was employed to obtain optimal coefficients. The entries defined and their proximity were the key factors to determine the optimum threshold and quality balance between such value as quality and quantity of results. As for expedient use, Machine learning techniques and Heavy NLP models (BERT, RNN) were considered for our research. With this, strong points and drawbacks of techniques examined stipulate the choice of the optimum one to solve the ED problem, conformably to our tasks and means.

The Use of BERT or RNN techniques demands a huge amount of various data, which will demand enormous resources for their collection and other procedures. Also, we have 2 levels of classes, for example, class *fear* with subclass *anxiety*. We can't drop subclasses, so we need to make second-level models, but again, the amount of data may be restricted and have a bias for specific conditions, so BERT or distilBERT training for multiclass classification doesn't even converge for both options. In addition, an essential factor for our research is the addition of new data, which must happen as quickly as possible. The supervised learning approach needs re-learning of our model to provide this function. It will take much more time compared to using an unsupervised learning model. Also, the speed of the

BERT models may be an issue when run on the CPU. So we are going back to embedding procedures and measuring the distance between them.

One of the key factors is embedding speed because we need to scan thousands of texts in a minute, not hours to make it available as soon as possible for analysts. The time of embedding creation for both models is provided in Table 5.

Table 5. Comparative indexes of using techniques for emotions detection

Task	fasttext (ours)	distilBERT	BERT
Create embedding	0,000534	0.02618	0.037545

To resolve the kind of problems stated in our research, unsupervised machine learning means have the following advantages. They are more applicable to analyze real-world problems due to the vague prospect of seeing what the outcomes should be and determining how accurate they are. Unsupervised machine learning purports to uncover previously unknown patterns in data, that corresponds to our task of unknown text processing. Unsupervised machine learning algorithms infer patterns from a dataset without reference to known or labeled, outcomes.

Besides, helpful unsupervised ML techniques use intakes the following:

- clustering, which allows splitting the dataset into groups automatically, according to similarity. Often, however, cluster analysis overestimates the similarity between groups and doesn't treat data points as individuals.
- association mining that identifies sets of items that frequently occur together in a dataset.
- latent variable models that are commonly used for data preprocessing, such as reducing the number of features in a dataset (dimensionality reduction) or decomposing the dataset into multiple components.

The patterns uncovered with unsupervised machine learning methods may also come in handy when implementing supervised machine learning methods later on. For example, an unsupervised technique might be used to perform cluster analysis on the data, then use the cluster to which each row belongs as an extra feature in the supervised learning model (semi-supervised machine learning) [23].

Unsupervised machine learning with The FastText embedding was chosen to solve the problem of class misbalance, to economize time and means, that substitutes training and use of supervised neural networks (BERT, RNN) approach in production.

The method is advisable for fast inference, both data processing and emotions prediction process without graphic processor unit (GPU) use.

As a result, we have obtained the two-modules system that encloses text processing and phrase emotion prediction. It provides the independent function of each of the parts (aka modules) in case of changes implemented.

The procedures set is contained in the standard procedures library of text processing and encloses patterns and exceptions to the rules. Such regulations are the tool to arrange a particular text and the whole text amount for processing.

In this way, all data obtained provide results of analysis interpretation, for Phase 3, with the following input data:

1. Processed text that is the result of the previous phases realization;
2. The table of alphanumeric codes and names of classes;
3. Procedures of synthesis of the output information.

Thus, the technique of text nature assessment described herein is the inherent part of our developed analytics approach to determine key text characteristics. The application prototype developed with this mean performs functions of Emotions detection with their thorough analysis which includes measuring of Emotional coloring of news texts and estimation of Intensity of news workflow impact.

The system prototype performs distribution of news by influence rating depending on their emotional coloring and news subject. It encloses two applications that work independently. They provide continuous news retrieval, downloads their texts, emotions, and subject detection, with saving the information. News pages are downloaded by headlines, news texts are separated from other content and are ranged by their information impact intensity. The example of text Emotional nature recognition, with Emotional coloring assessment for mass media news texts [24 – 28] are presented hereinafter (Fig.1 – Fig.5). As it is seen, this characteristic is defined properly, with prevalent emotions corresponding to the most of the news content.

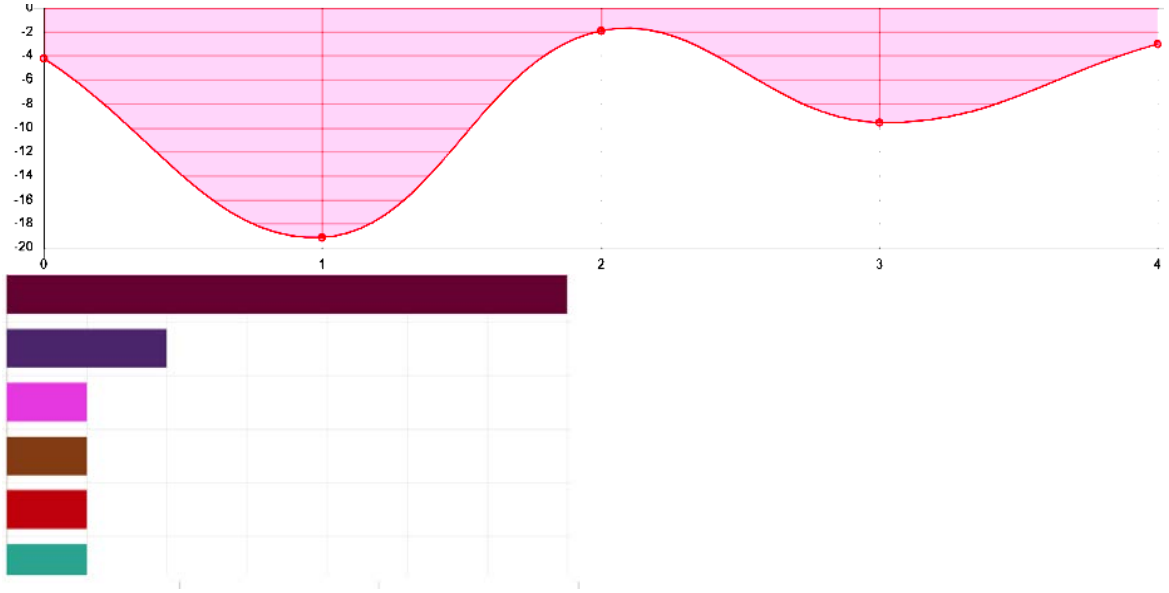


Fig. 1. The visualization of results of mass media news text emotional coloring assessment for the news 1, “Ammonium nitrate: “What is it and why did it cause the blast in Beirut?”

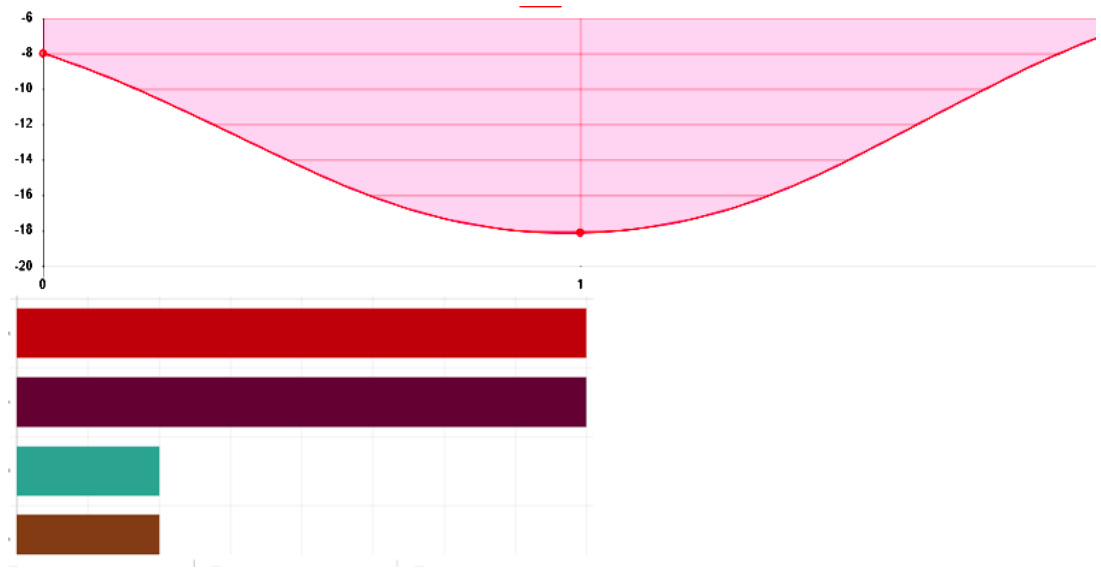
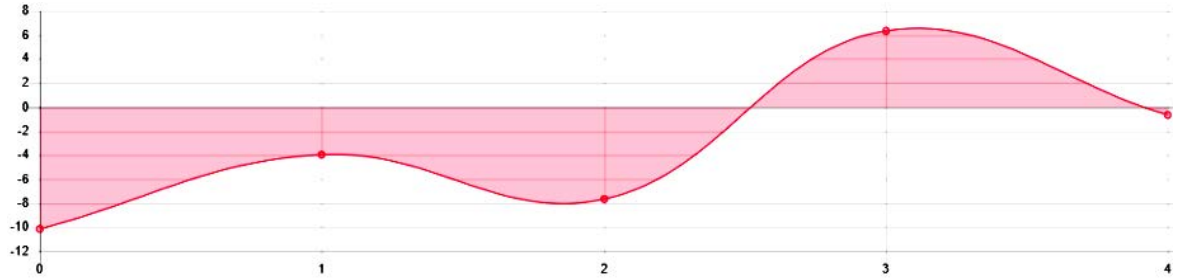


Fig.2. The visualization of results of mass media news text emotional coloring assessment, news 2, “SBU detains member of notorious Vostok Battalion involved in Donetsk airport battles”.



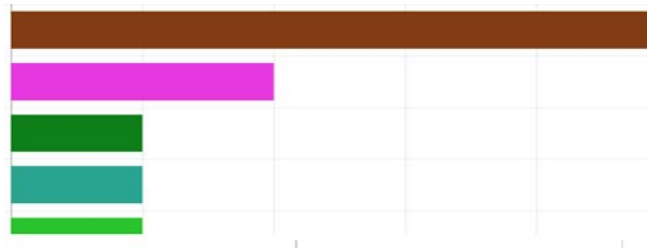


Fig.3. The visualization of results of mass media news text emotional coloring assessment, news 3 “How the world is coping with coronavirus, six months on”.

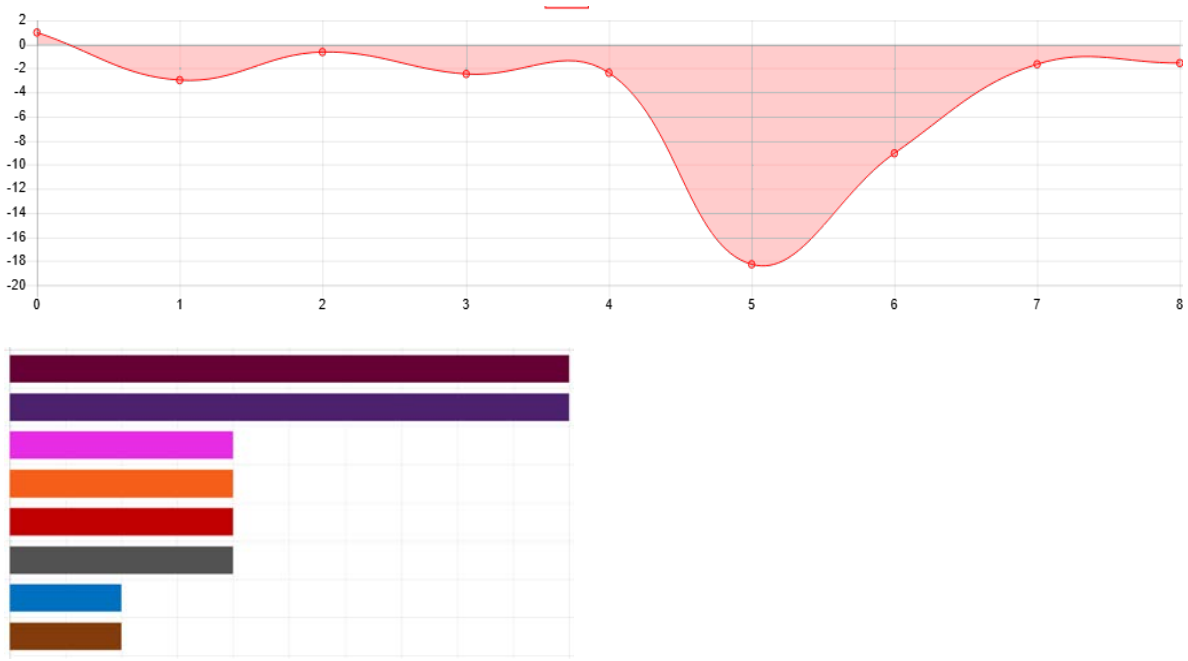


Fig.4. The visualization of results of mass media news text emotional coloring assessment, news 4, “Reports of 40 Chinese casualties in border clash with India are ‘fake news’ – Chinese Foreign Ministry”.

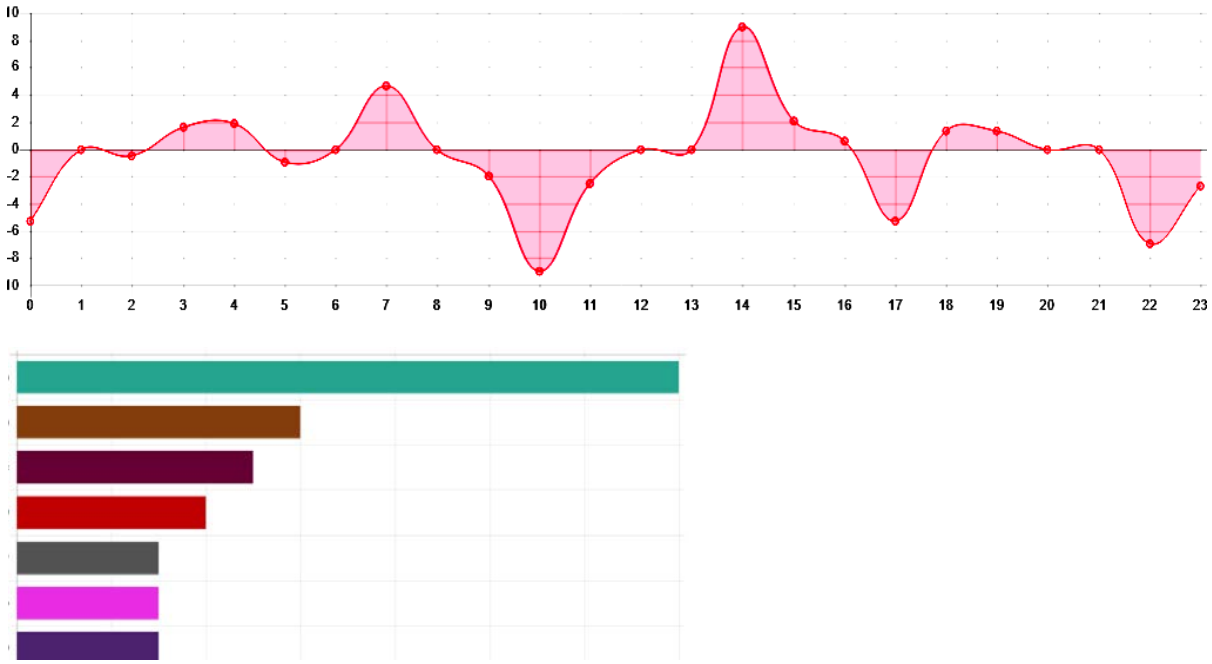


Fig.5 The visualization of results of mass media news text emotional coloring assessment, news 5, “Merger of Dfid and Foreign office risks lack of transparency over 14 bn£ aid budget”.

The visualization of results includes marking of Emotional coloring of each news text by the color appropriate to the particular emotion. The diagram for each text shows a correlation of emotions that are inherent to the particular news text.

On each graph above-seen the axis X corresponds to the progress of the text and the number of its figures is equal to the number of sentences that form each news. The axis Y means the intensity of Emotional coloring for each news text, with compliance of its degrees with neutral coloring (as for 0 marks) and positive or negative Emotional coloring for fields below and above 0 marks, in the positive or negative field correspondingly.

The diagram beneath each graph demonstrates the presence of certain emotions in the news text and their correlation according to their share in the text. Each emotion is signed with a particular color, so it is easy to see at a single glance the nature of the text. Depending on an analytical task, it is possible to visualize the nature of the entire informational workflow processed for a certain period.

5. Conclusions

The results of research for the emotions detection technique stated in this article present a new processing model which complies with the needs for Mass media processing in a turbulent political environment. The procedure provides a thorough evaluation of emotional characteristics of Mass media news texts, with measurement of their degree of intensity and assessment of the entire information workflow. Rapid processing of data of defined format and multilateral information is realized and the addition of new data of defined format replying to political and social environment changes is supported.

The newly developed technique encloses the means for key text characteristics analysis. They form the set of rules and procedures, which apply to the text Emotional coloring assessment. Machine learning tools are the key means that make possible an appropriate processing and identification of text Emotional coloring.

A formal description is based on diverse factors that influence the power of the information workflow impact and form the Emotional coloring of the news text. Its Emotional coloring is defined by Emotions intensity characteristics, which are inherent to the text and defined with alphanumerical code obtained with the expert evaluation procedure.

Wordforms structure recognition is provided through Lexical-semantic rules that recognize sentence semantic categories and provide identification of wordforms structures.

Logical-semantic rules that regulate wordforms interaction. They are the main regulation to define the proper weighting ratio for emotionally colored wordforms, which are an indication of such coloring for their sentences and the whole text. Moreover, these rules are an integral part of data formal presentation for key text characteristics analysis.

Implementation of the means proposed contributes to the text's adequate apprehension and facilitates recognition of its true emotional nature.

The Classification rules that enable the identification of word structures by syntactic and semantic categories is an important means to regulate the identification of emotions through wordforms for the news text.

The efficiency of these techniques is supported with the most adequate models, SVM and SVM with SGD that provide the optimum indexes for a new system prototype functioning. The toolbox (FastText embedding and Elastic Search) assures an appropriate rate and precision of text processing within an unsupervised machine learning method. They are helpful to explore a real-world environment with a high degree of uncertainty and new unregulated information.

The Prospect of our research regards both linguistic and psychological aspects of Mass media study. Including assessment of polarity of an audience, which might be an object for improvement, for better results in Sentiment analysis procedures, concerning the audience's positive and negative attitudes.

Drawbacks of the described technique concern its operational abilities that are usable for Mass media sites and other resources where text content has a distinct structure. Social networks are not a subject for processing now. For the present, our model is designed for texts with a strict grammar structure and syntactic order, which are appropriate for official Mass media. As for social network texts, their content is often unstructured and chaotic, without a distinct structure, abundant of slang, and explicitly modified lexical units.

Besides, now the system disposes only of two operational languages, English and Russian, with a set of rules and procedures and related linguistic means corresponding to them exactly.

A strong point of the presented technique is the following: a multilateral assessment of Mass media news content, with a focus on the degree of intensity of emotions manifestation in the text, Emotional coloring and an overall Intensity of information impact of a text. The newly-developed technique for ED and other text characteristic analysis affords to identify an entire syntactic structure inside of a text and to measure its parameters, which distinguishes it from other antecedent ED means. All this enables a thorough assessment of mass media news texts both for their emotional nature and multilateral analysis procedure, which is accompanied by practical visualization assets. Its employment in the interests of Mass media news analysis makes it possible to see particular details of news content in key text characteristics that mirror a turbulent social and political environment.

References

[1] King1,G., Schneer,B., White A. (2017). How the News Media Activate Public Expression and Influence National Agendas. [Online]. Available: https://gking.harvard.edu/files/gking/files/776.full_.pdf (accessed 23 September 2021)

[2] Kanischeva, O., Medvedska, A., Panchul, O. (2014).Vyznachennia Typiv Emotsiynogo Movnogo Vyslovlyuvannia u Dodatkah Avtomatychnogo Opratsyuvannia Tekstiv. [Online]. Available: http://science.lp.edu.ua/sites/default/files/Papers/31_119.pdf (accessed 23 September 2021)

[3] Pasquier C., da Costa Pereira C., Tettamanzi, A. G. B. (2020). Extending a Fuzzy Polarity Propagation Method for Multi-Domain Sentiment Analysis with Word Embedding and PosTagging. [Online]. Available: https://www.semanticscholar.org/paper/Extending-a-Fuzzy-Polarity-Propagation-Method-for-Pasquier-Pereira/32b87b4ab00e2bc3f2 bbcda1dd946a8d405980c6 (accessed 23 September 2021)

[4] Valdivia, A., Luzon, M. V., Herrera, F. (2017). Sentiment Analysis in TripAdvisor. [Online]. Available: https://www.computer.org/publications/tech-news/research/tripadvisor-algorithm-sentiment-analysis-tourism-research (accessed 23 September 2021)

[5] Schnoll, M., Ferner, C., Wegenkittl, S. (2019). The Effectiveness of the Max Entropy Classifier for Feature Selection. [Online]. Available: https://www.researchgate.net/publication/336907526_The_Effectiveness_of_the_Max_Entropy_Classifier_for_Feature_Selection (accessed 23 September 2021)

[6] Hu, Z. Mukhin, V. Kornaga, Y. Lavrenko, Y. Herasymenko, O. "Distributed computer system resources control mechanism based on network-centric approach". International Journal of Intelligent Systems and Applications, 2017, 9(7), pp. 41-51.

[7] Z. Hu, V. Mukhin, Ya. Kornaga, O. Herasymenko and Ye. Mostoviy. "The Analytical Model for Distributed Computer System Parameters Control Based on Multi-factoring Estimations". Journal of Network and Systems Management, vol. 27, no. 2, pp. 351-365, 2019.

[8] Mukhin, V., Volokyta, A., Heriatovych, Y., Rehida, P. "Method for efficiency increasing of distributed classification of the images based on the proactive parallel computing approach". Advances in Electrical and Computer Engineering, 2018, 18(2), pp. 117–122.

[9] Zhengbing, H., Mukhin, V.Y., Kornaga, Y.I., Herasymenko, O.Y. "Resource Management in a Distributed Computer System with Allowance for the Level of Trust to Computational Components". Cybernetics and Systems Analysis, 53 (2), pp. 312-322. doi: 10.1007/s10559-017-9931-9

[10] Mukhin V., Kuchuk N., Kosenko N., Artiukh A., Yelizyeva A., Maleyeva O., Kuchuk H., Kosenko V. Decomposition Method for Synthesizing the Computer System Architecture. Advances in Computer Science for Engineering and Education II. ICCSEEA 2019. Advances in Intelligent Systems and Computing, vol 938. Springer, Cham. https://doi.org/10.1007/978-3-030-16621-2_27

[11] Alexander Dodonov, Vadym Mukhin, Valerii Zavgorodnii, Yaroslav Kornaga, Anna Zavgorodnya, Oleg Mukhin, "Method of Parallel Information Object Search in Unified Information Spaces", International Journal of Computer Network and Information Security, Vol.13, No.4, pp.1-13, 2021.

[12] Roberts, H., Resch, B., Sadler, J., Chapman, L. Petutschnig, A., Zimmer, S. (2018). Investigating the Emotional Responses of Individuals to Urban Green Space Using Twitter Data: a Critical Comparison of Three Different Methods of Sentiment Analysis. Urban Planning. https://www.researchgate.net/publication/324118687_Investigating_the_Emotional_Responses_of_Individuals_to_Urban_Green_Space_Using_Twitter_Data_A_Critical_Comparison_of_Three_Different_Methods_of_Sentiment_Analysis (accessed 23 September 2021)

[13] Acheampong, FA, Wenyu, C, Nunoo-Mensah, H. (2020). Text-Based Emotion Detection: Advances, Challenges, and Opportunities. Engineering Reports; 2:e12189. [Online]. Available: https://doi.org/10.1002/eng2.12189 (accessed 23 September 2021).

[14] Hasan, M., Elke, A., Rundensteiner, E.Agu. (2014). EMOTEX: Detecting Emotions in Twitter Messages. [Online]. Available: https://web.cs.wpi.edu/~emmanuel/publications/PDFs/C30.pdf (accessed 23 September 2021).

[15] Izard, K. (2012). Psihologiya emocij. – Spb.: Piter,. — 464 p.

[16] Babich, O., Popov, N., Glukhov, S. "The basics for development of mass media information stream classifier" Proceedings of AC 2019 in Prague. (8-10.08.2019). Czech Technical University in Prague. P.157–164.

[17] Rodney Huddleston. A short overview of English syntax. The University of Queensland. [Online]. Available: http://www.lel.ed.ac.uk/grammar/overview.html (accessed 23 September 2021).

[18] Cristianini N., Ricci, E. (2008) Support Vector Machines. In: Kao MY. (eds) Encyclopedia of Algorithms. Springer, Boston, MA. [Online]. Available: https://doi.org/10.1007/978-0-387-30162-4_415 (accessed 23 September 2021).

[19] Gentile, C, Warmuth, M. (1998). Linear Hinge Loss and Average Margin. [Online]. Available: https://www.researchgate.net/publication/220270147_Linear_Hinge_Loss_and_Average_Margin (accessed 23 September 2021).

[20] Srikumar, V. (2018). Support Vector Machines: Training with Stochastic Gradient Descent. Machine Learning. [Online]. Available: https://www.cs.utah.edu/~zhe/pdf/lec-19-2-svm-sgd-upload.pdf (accessed 23 September 2021).

[21] Bojanowski, P., Grave, E., Joulin, A., Mikolov, T. (2017). Enriching Word Vectors with Subword Information. [Online]. Available: https://arxiv.org/pdf/1607.04606.pdf) (accessed 23 September 2021).

[22] Malkov, Yu., Yashunin, D. Efficient and Robust Approximate Nearest Neighbor Search Using Hierarchical Navigable Small World Graphs. [Online]. Available: https://arxiv.org/pdf/1603.09320.pdf (accessed 23 September 2021).

[23] Yang, X, Song, Z, King, I., Xu, Z. (2021). A Survey on Deep Semi-supervised Learning https://arxiv.org/pdf/2103.00550v2.pdf (accessed 23 September 2021).

[24] The Telegraph. (2020). Ammonium nitrate: what is it and why did it cause the blast in Beirut? [Online]. Available: https://www.telegraph.co.uk/news/2020/08/06/beirut-explosions-ammonium-nitrate/?ICID=escenic-liftigniter_recommentation-widget&li_source=LI&li_medium=escenic-section (accessed 22 September 2021)

[25] The UNIAN, Information agency. (2020). "SBU detains member of notorious Vostok Battalion involved in Donetsk airport battles". [Online]. Available: https://www.unian.info/war/donbas-war-sbu-detains-vostok-battalion-member-involved-in-donetsk-airport-battles-11102555.html (accessed 22 September 2021)
[26] The Guardian. (2020). How the world is coping with coronavirus, six months on. [Online]. Available: https://www.theguardian.com/news/audio/2020/aug/05/how-the-world-is-coping-with-coronavirus-six-months-on (accessed 22 September 2021)
[27] The Russia today, Information agency. (2020). Reports of 40 Chinese casualties in border clash with India are 'fake news' – Chinese Foreign Ministry. [Online]. Available: https://www.rt.com/news/492661-china-india-reports-casualties-fake/ (accessed 22 September 2021)
[28] The Telegraph. Merger of Dfid and Foreign office risks lack of transparency over 14 bn£ aid budget. (2020). [Online]. Available: https://www.telegraph.co.uk/global-health/climate-and-people/merger-dfid-foreign-office-risks-lack-transparency-14bn-aid/ (accessed 22 September 2021).

Authors' Profiles

Oksana Babich: Senior researcher of the Research center of the Taras Shevchenko National University, Kyiv, Ukraine.

The Specialist degree in foreign linguistics, The Taras Shevchenko National University, Kyiv (2002), The Specialist degree in Project Management, the State University of Telecommunications (2017).

PhD in technics (2011), from The Taras Shevchenko National University, Kyiv.

Major interest: psycholinguistics, information and analytical activity, machine learning.

Viktor Vyshnivskiy: Professor of department of Computer science of the State University of Telecommunications; Doct. of Sc., Professor

PhD in technics (1994), from the Military Institute of Communication of Kiev, Professor (2013), Doct. of Sc. (2016), from the State University of Telecommunications.

Major interests: information technologies, information and cybernetics security, information systems diagnostics and security.

Vadym Mukhin: Professor of the department of Mathematical methods of system analysis, the National Technical University of Ukraine "Kiev Polytechnic Institute", Doct. of Sc.

PhD (1997), Doct. of Sc. (2015) from the National Technical University of Ukraine "Kiev Polytechnic Institute"; Professor (2015).

Major interests: the security of distributed computer systems and risk analysis; design of the information security systems; mechanisms for the adaptive security control in distributed computing systems; the security policy development for the computer systems and networks.

Irina Zamaruyeva: Professor of department of Computer science of the State University of Telecommunications; Doct. of Sc., Professor.

Major interest: information technologies, machine learning, information and cybernetics security, information and analytical activity, psycholinguistics.

Michail Sheleg: a student of the computer science department for Bachelor degree, the State University of Telecommunications, Kyiv, Ukraine.

Software engineer (Machine Learning, backend) and project manager.

Major interest: machine learning and data analysis, natural language processing, computer vision.

Yaroslav Kornaga: Assoc. professor of Computer systems department of National Technical University of Ukraine “Kiev Polytechnic Institute”, PhD.

(2015), Doct. of Sc. (2020) from the State University of Telecommunications; Assoc. Prof. (2015) of technical cybernetics department.

Major interests: the security of distributed database and risk analysis; design of the distributed database; mechanisms for the adaptive security control in distributed database; the security policy development for distributed database.

How to cite this paper: Oksana Babich, Viktor Vyshnyvskiy, Vadym Mukhin, Irina Zamaruyeva, Michail Sheleg, Yaroslav Kornaga, "The Technique of Key Text Characteristics Analysis for Mass Media Text Nature Assessment", International Journal of Modern Education and Computer Science(IJMECS), Vol.14, No.1, pp. 1-16, 2022.DOI: 10.5815/ijmecs.2022.01.01