# Applying VSM to Identify the Criminal Meaning of Texts

Nina Khairova[1] [0000-0002-9826-0286], Anastasiia Kolesnyk[1] [0000-0001-5817-0844], Orken Mamyrbayev[2] [0000-0001-8318-3794] and Svitlana Petrasova[1] [0000-0001-6011-135X]

[1]National Technical University "Kharkiv Polytechnic Institute", 2, Kyrpychova str., 61002, Kharkiv, Ukraine
[2]Institute of Information and Computational Technologies, 125, Pushkin str., 050010, Almaty, Republic of Kazakhstan
khairova@kpi.kharkov.ua, kolesniknastya20@gmail.com, morkenj@mail.ru, svetapetrasova@gmail.com

**Abstract.** Generally, to define the belonging of a text to a specific theme or domain, we can use approaches to text classification. However, the task becomes more complicated when there is no train corpus, in which the set of classes and the set of documents belonged to these classes are predetermined. We suggest using the semantic similarity of texts to determine their belonging to a specific domain. Our train corpus includes news articles containing criminal information. In order to define whether the theme of input documents is close to the theme of the train corpus, we propose to calculate the cosine similarity between documents of the corpus and the input document. We have empirically established the average value of the cosine similarity coefficient, in which the document can be attributed to the highly specialized documents containing criminal information. We evaluate our approach on the test corpus of articles from the news sites of Kharkiv. F-measure of the document classification with criminal information achieves 96 %.

**Keywords:** semantic similarity of texts, VSM, criminal information, news sites, cosine similarity, PPMI

## 1 Introduction

One of the main tasks of NLP and, accordingly, of computer linguistics, in general, is the task of semantic similarity of different elements of the texts (words, phrases, colocations, sentences and documents). This task is directly related to information retrieval, ranking of documents, topic modeling of texts, sentiment analysis and more.

The task of the identification of the documents semantic similarity is used in all approaches that utilize semantic analysis and semantic technologies including the monitoring of public information, telecommunication networks. However, mostly, this task, which originally was considered by Salton [1], is applied regarding information retrieval. In considering the issue of the documents similarity identification, Salton et. al (1975) focused on measuring document similarity, considering a query to search engine as a pseudo-document.

Copyright © 2020 for this paper by its authors.
Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR-WS.org/Vol-2604/paper2.pdf

We suggest using the semantic similarity of texts to determine their belonging to a specific domain. Usually, the solutions of the task are based on methods and approaches to text classification. Good and frequently used methods of text classification are the decision tree, neural networks [2], Random Forest and Support Vector Machine [3], the Bayesian method, K-means [4] and others like them [5]. Nevertheless, all these methods require a trained corpus, in which the set of classes and the set of documents belonged to these classes, are predetermined.

In our case there are no predefined classes, we have only a text corpus of a specific domain, which includes news articles containing criminal information. It the study, in order to determine belonging of a text to the specific domain, when we cannot use classification because there are no predefined classes, we suggest measuring the similarity of the texts.

The remainder of the paper is organized as follows. Section 2 gives an overview of the related works, corresponding with methods and approaches of semantic similarity of texts. Section 3 describes the application of VSM for semantic analysis. Section 4 presents the usage of our method for identifying the criminal meaning of texts. Section 5 introduces our corpus comprising texts contained criminal information and describes its usage in our experiment. In the last Section 6, the scientific and practical contributions of the research, its limitations and future work are discussed.

## 2 Related Work

Search for the semantic similarity of text information is getting more and more popular in various fields, for instance, [6] utilized semantic similarity as a very effective method to identify links between medical objects such as a drug and a diagnosis. Their approach is based on the transformation of embedding into a drug-prescription model and assesses similarities between them using a vector representation of the link between drug and prescription. This approach has been empirically studied and shows good results in bio-medicine.

Evaluation of semantic similarity of texts also helps in market analysis, banking and marketing. In paper [7] authors used this approach to determine the similarities between various press releases of a bank and to assess their impact on potential clients and the financial market. This was done by calculating the distance between fixed- vectors length of pairs of press releases (bag of words model). The method assigned more weight to words that were rare and less weight to words that were frequent. Testing also showed that the results were not sensitive to weighting.

This technique calculated the semantic similarity between the text words and the lexical dictionary is also exploited in the field of sentiment analysis, especially for processing various lexical resources. The authors of the study [8] applied this metric for sentiment classification model using a measure of semantic proximity and embedding representations. However, the results of the study indicate that the choice of the vocabulary influences cross-dataset analysis.

The study [9] used a new method to determine the semantic similarity of big documents that were academic articles. These articles had some kind of topic events that

presented the same information about research objectives, methodologies, etc., as well. To calculate the degree of semantic similarity, the authors of the study exploited domain ontology, which was used to calculate similarities between semantic events. Estimated experiments had shown that such a method based on ontologies got more accurate results compared to others. Generally, the methods, which utilized ontologies, showed overall advantages in Correlation, Accuracy and F1-score.

Overall, a lot of studies provided different ways to improve the computation of semantic similarity. For instance, study [10] provided good results for the new vector space model based on a random walk algorithm. The peculiarity of this approach was the comparison of the distribution of each text that induces when used as the seed of a random walk over a graph derived from WordNet. This algorithm had a relative decrease in the error rate in comparison to a conventional vector model.

Traditionally, there are two groups of approaches to the task of semantic similarity identification. The first group is based on ontologies. For instance, the ontology-based approach was utilized by Resnik's method [11] or the extended Lesk Algorithm [12]. However, in the vast majority of cases, such approaches applied to identify the semantic similarity of short text fragments or words.

The second group of approaches to the task of semantic similarity identification is based on statistical methods of distributional similarity. These methods applied to measure words similarity as well as to similarity of documents and even similarity of relations [13]. Therefore, the semantic similarity of large documents is still based only on statistical information, which is clearly insufficient for determining global semantic values [14].

## 3 The Application of VSM for Semantic Analysis

The purpose of our study is to find a universal method for solving the problem of identification of the texts with criminal meaning. When determining the thematic orientation of texts, namely for the task of classification or clustering, vector space model is an adequate and well-developed method.

The vector space model (VSM) allows providing a collection of documents by vectors from one vector space, which is common for the whole collection of documents [13]. The use of VSM is based on two cognitive hypotheses. The first hypothesis, statistical semantics hypothesis, states that statistical patterns of word usage in natural language can be used to find out what people mean. In other words, human intellect can understand words depending on their environment [15].

The second hypothesis, formulated by J. Salton for information retrieval [1], is based on the representation of the text as "a bag of words" and suggests that the frequency of words in a document often determines the relevance of documents to the query.

The main idea of VSM is to represent each collection document as a point in multidimensional space (vector in vector space). The points lying close to each other correspond to semantically similar documents. Therefore, in the vector space model, text representation mainly focuses on two tasks. Firstly, how to build a vector and, secondly, how to assign weights to vector elements.

Towards the first objective, each document in a vector model is thought as an unordered set of terms. The terms can be any words, including numbers and proper names. With a large collection of researched documents, as in our case, and correspondingly a large number of vectors, it is reasonable to place the data in the matrix. Each row of the matrix defines a separate term, and each column corresponds to some document.

The second primary task of VSM is to determine the weight of the terms in the document. Weight refers to the importance of a word, its semantic ability to identify a given text. The easiest way is to determine the frequency of a term.

Nevertheless as usual, in order to determine the weight of a term in a term-document matrix, tf*idf index is used, which stands for "term frequency * inverse document frequency".

The purpose of weighing the terms is to determine how fully they reflect the semantic content of the document. However, frequency and probabilistic methods of tf*idf have a number of disadvantages, as often the result may be irrelevant documents or lack of true relevance. Such problems with the result are related to the fact that the methods described do not take into account that the frequencies of occurrence of different terms depend on each other, since they can be combined into word combinations. In addition, it affects the result and the synonymy and plurality of the language.

In order to solve some of these problems we will use PMI (Pointwise Mutual Information) as a weight function [16]. Formally, PMI can be defined in the following way.

Let $F$ be a traditional term-document matrix in which $n_r$ rows and $n_c$ columns, i-$_{th}$ row in matrix $F$ is the vector of row $f_{i:}$ and j-$_{th}$ column in matrix $F$ is the vector of column $f_{:j}$. Row $f_{i:}$ corresponds to term $w_i$ and column $f_{:j}$ corresponds to document $d_j$. The value of element $f_{ij}$ is the number of times that $w_i$ appears in document $d_j$.

Let $X$ be the matrix that is obtained by calculating the PMI weight function to the elements of matrix $F$. Then matrix $X$ will have as many rows and columns as the frequency matrix $F$. The values of the $x_{ij}$ element in matrix $X$ are defined as follows equations:

$$p_{ij} = f_{ij} / \sum_{i=1}^{n_r} \sum_{j=1}^{n_c} f_{ij} \tag{1}$$

$$p_{i*} = \sum_{j=1}^{n_c} f_{ij} / \sum_{i=1}^{n_r} \sum_{j=1}^{n_c} f_{ij} \tag{2}$$

$$p_{*j} = \sum_{i=1}^{n_r} f_{ij} / \sum_{i=1}^{n_r} \sum_{j=1}^{n_c} f_{ij} \tag{3}$$

where:

$p_{ij}$ is the estimated probability that term $w_i$ will appear in document $d_j$;

$p_{i*}$ is the estimated probability of term $w_i$, i.e. the probability that the term will appear in any collection document;

$p_{*j}$ is the estimated probability of document $d_j$, i.e probability that the document will appear with any term.

To determine PMI, we calculate the logarithm of the estimated probability $p_{ij}$ (1) divided by the product of two probabilities $p_{i*}$ (2) and $p_{*j}$ (3):

$$pmi_{ij} = \log(p_{ij} / (p_{i*} p_{*j})), \quad (4)$$

If $w_i$, and $d_j$ are statistically independent, then according to the definition of the product of probability of independent events, $p_{i*} * p_{*j} = p_{ij}$. In this case, the value of the formula logarithm of the definition of $pmi_{ij}$ (4) will be zero. However, if there is some semantic interrelation between $w_i$, and $d_j$, it should be expected that $p_{ij}$ will be more than it would be if $w_i$, and $d_j$ were semantically independent. Therefore, we should look for the PPMI (Positive Pointwise Mutual Information) weight function, which is defined as:

$$x_{ij} = ppmi_{ij} = \begin{cases} pmi_{ij}, \text{ if } pmi_{ij} > 0 \\ 0, \text{otherwise} \end{cases}, \quad (5)$$

To measure the similarity of two weighted frequency vectors, we will determine their cosine similarity [17]. Let $x$ and $y$ be two vectors of $n$ elements. Then the cosine of angle $\Theta$ between vectors x and y can be calculated as inner product of vectors normalized by their lengths.

$$\cos(x, y) = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i^2}} = \quad (6)$$

To calculate the cosine between vectors $x$ and $y$, we summarize the products of their coordinates $x_1 \bullet y_1 + x_2 \bullet y_2 + \ldots + x_n \bullet y_n$, and then divide this product into a square root of the sum of squares of their coordinates (6).

According to formula (6), the value of a cosine can vary from minus one if vectors have opposite directions ($\Theta$ =180 degrees) to plus one if directions of vectors coincide ($\Theta$ =0 degrees). When vectors are perpendicular ($\Theta$ =90 degrees), the cosine of the angle between them is equal to zero. Since by definition of PPMI weights cannot be negative, therefore cosine values between vectors that use PPMI as coordinates will always lie in the positive range [0, 1].

## 4 Our Method for Identifying the Criminal Meaning of Texts

The main objective that we are seeking to achieve as a result of this research is to find a universal method for determining the semantic similarity of the input text to a particular thematic focus. Namely, we define the thematic closeness to texts that contain criminal information.

To determine whether a random text belongs to a criminal subject, we use self-created corpus containing news articles related to criminal content. In fact, the concept of "the criminal meaning of texts" is blurry and subjective. We put the following meaning into it: the text belongs to the category of "Criminal" if it contains information about emergency news, war, terrorism, accidents, extremism, criminal offences, etc.

Within this topic, the corpus of criminal texts and the corpus with the materials, which do not correspond to the investigated subject, were collected.

Our method for determining whether a document belongs to a highly specialized area includes three main steps, as shown in Figure 1: (1) linguistic processing of the

raw corpus and the input document; (2) machine learning phase; (3) cosine similarity phase.

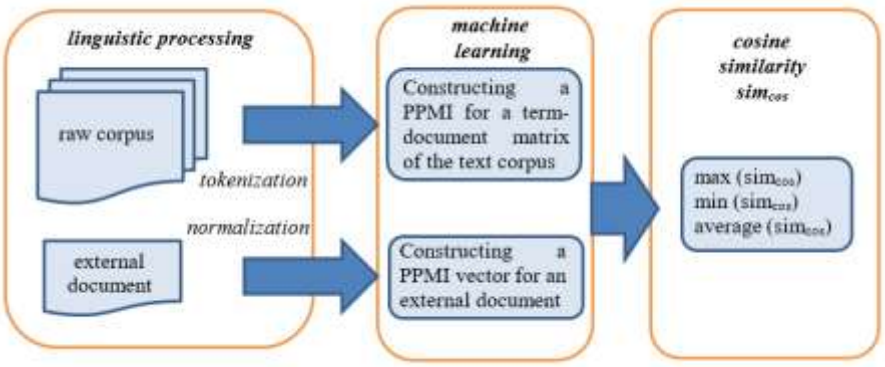


**Fig. 1.** The general scheme of the applied method.

At the first stage of creating the term-document matrix the linguistic processing of raw texts consisting of tokenization and text normalization was performed.

The second stage of processing is the stage of machine learning, which consists of building a PPMI term-document matrix and PPMI vector of the input document. This stage is based on the method described in the previous subsection.

At the third stage, the minimum, maximum and average values of the semantic similarity coefficient were calculated, defined in cosine similarity between the input document vector and the document vectors of the available PPMI term-document matrix.

Figure 2 shows an example of the developed application MainWindow, which allows identifying the criminal meaning of the incoming news text.

## 5 Description of the Corpus and Result of the Experiment

Our dataset includes the corpus of criminal texts and the corpus with articles, which do not correspond to this topic. To create the first corpus named "*criminal*" articles were taken from the news sites of Kharkiv, such as: 057.ua, Mediaport, ATN, Vechirniy Kharkiv, Misto by the categories of war, terrorism, accidents, extremism, criminal offences, etc. in the period 2007-2018 [18]. For the second corpus named "*non-criminal*" texts were collected automatically from the same information sites, but in other categories. In general, the volume of these two corpora is more than 195,000 text files in *text* format.

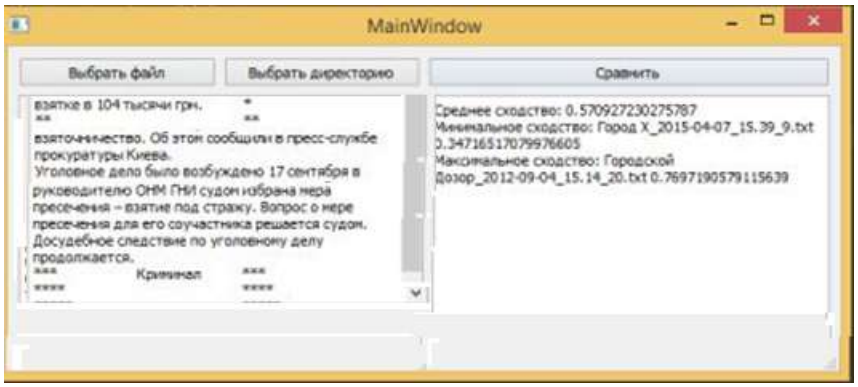


**Fig. 2.** The example of the running program MainWindow

In order to determine the value of the coefficient $sim_{cos}$ allowing attributing the input document to the criminal theme, we made the following experiment. We have allocated 90% of texts of the "*criminal*" corpus into a train corpus. The test corpus has consisted of the last part of texts from "*criminal*" corpus and texts from "*non-criminal*".

During the experiment, the values of cosine similarity of tested text $doc_{test}$ from "criminal" part of the test corpus and each text from the trained corpus $sim_{cos}$ ($doc_{test}$, $doc_{train}$) was calculated. Table 1 shows the maximum, minimum and average values of cosine similarity between each "criminal" text of the test corpus and documents of the train corpus, which also belong to the criminal theme.

**Table 1.** The fragment of the analysis results of $sim_{cos}$ between criminal texts of the test corpus and criminal texts of the train corpus

| File | Cosine similarity $sim_{cos}$ | | |
|---|---|---|---|
| | Max ($sim_{cos}$) | Min ($sim_{cos}$) | Average ($sim_{cos}$) |
| ATN_2018-08-31_11.33_1.txt | 0,92 | 0,26 | 0,72 |
| ATN_2018-08-31_12.27_2.txt | 0,77 | 0,35 | 0,57 |
| ATN_2018-08-31_12.33_3.txt | 0,79 | 0,38 | 0,67 |
| ATN_2018-08-31_16.24_4.txt | 0,78 | 0,38 | 0,67 |
| Misto X_2018-08-31_16.39_5.txt | 0,82 | 0,36 | 0,69 |
| Misto X_2018-09-01_17.44_6.txt | 0,80 | 0,39 | 0,68 |
| Misto X_2018-09-02_10.26_7.txt | 0,88 | 0,27 | 0,69 |
| Misto X_2018-09-02_12.58_8.txt | 0,80 | 0,27 | 0,60 |
| Misto X_2018-09-02_17.00_9.txt | 0,87 | 0,31 | 0,71 |
| VKh_2018-01-14_13.50_2.txt | 0,77 | 0,50 | 0,66 |
| VKh _2018-01-14_14.10_3.txt | 0,82 | 0,48 | 0,68 |
| VKh_2018-01-14_14.40_4.txt | 0,77 | 0,49 | 0,67 |
| Mediaport _2018-01-15_15.10_5.txt | 0,73 | 0,28 | 0,51 |

| | | | |
|---|---|---|---|
| Mediaport _2018-01-16_11.10_6.txt | 0,81 | 0,36 | 0,68 |
| Mediaport _2018-01-16_15.00_7.txt | 0,88 | 0,31 | 0,71 |
| Mediaport _2018-01-16_18.40_8.txt | 0,85 | 0,33 | 0,69 |
| 057.ua _2018-10 16_00.09_1.txt | 0,79 | 0,42 | 0,70 |
| 057.ua _2018-10-17_00.13_2.txt | 0,67 | 0,41 | 0,59 |
| 057.ua _2018-10-17_00.18_3.txt | 0,89 | 0,37 | 0,67 |
| 057.ua _2018-10-18_00.18_4.txt | 0,74 | 0,38 | 0,61 |
| 057.ua _2018-10-19_00.02_5.txt | 0,77 | 0,43 | 0,62 |
| 057.ua _2018-10-22_00.19_6.txt | 0,84 | 0,43 | 0,72 |

The analysis of the obtained results allowed concluding that the minimum value of cosine similarity coefficient ($sim_{cos}$) of the texts containing criminal information of the test corpus and documents of the trained corpus is not less than 0.3 ($min(sim_{cos})> 0.3$), the maximum value of $max(sim_{cos})> 0.7$, and the average value of $average(sim_{cos})> 0.55$.

Table 2 shows certain values of *min* ($sim_{cos}$), *max* ($sim_{cos}$) and *average* ($sim_{cos}$) of cosine similarity between documents of the train corpus and documents from the test corpus, which can be related to any subject except for the criminally specialized one.

**Table 2.** The fragment of the $sim_{com}$ results for the test corpus of a random theme

| File | Cosine similarity, $sim_{cos}$ | | |
|---|---|---|---|
| | Max ($sim_{cos}$) | Min ($sim_{cos}$) | Average ($sim_{cos}$) |
| ATN_2018-08-31_1.txt | 0,63 | 0,24 | 0,41 |
| ATN_2018-08-31_2.txt | 0,66 | 0,25 | 0,47 |
| ATN_2018-08-31_3.txt | 0,69 | 0,19 | 0,45 |
| ATN_2018-08-31_4.txt | 0,67 | 0,21 | 0,41 |
| ATN_2018-08-31_5.txt | 0,73 | 0,24 | 0,48 |
| ATN_2018-09-6.txt | 0,60 | 0,25 | 0,44 |
| Misto X_2018-09-7.txt | 0,51 | 0,15 | 0,36 |
| Misto X _2018-09-8.txt | 0,74 | 0,20 | 0,53 |
| Misto X _2018-09-9.txt | 0,67 | 0,22 | 0,47 |
| Misto X _2018-09-10.txt | 0,62 | 0,23 | 0,43 |
| Misto X _2018-09-11.txt | 0,76 | 0,19 | 0,55 |
| Misto X _2018-09-12.txt | 0,75 | 0,22 | 0,60 |
| Misto X _2018-10-19_13.txt | 0,72 | 0,25 | 0,51 |
| Mediaport_2018-10-19_14.txt | 0,65 | 0,19 | 0,47 |
| Mediaport _2018-10-19_15.txt | 0,66 | 0,19 | 0,50 |
| Mediaport _2018-10-19_16.txt | 0,62 | 0,21 | 0,38 |
| Mediaport _2018-10-19_17.txt | 0,56 | 0,28 | 0,45 |
| Mediaport _2018-10-19_18.txt | 0,62 | 0,20 | 0,44 |
| Mediaport _2018-10-19_19.txt | 0,60 | 0,18 | 0,39 |
| VKh_2018-10-19_20.txt | 0,65 | 0,24 | 0,43 |
| VKh _2018-10-19_21.txt | 0,73 | 0,25 | 0,50 |
| VKh _2018-10-19_22.txt | 0,65 | 0,18 | 0,44 |

| | | | |
|---|---|---|---|
| 057.ua _2018-10-19_23.txt | 0,59 | 0,23 | 0,47 |
| 057.ua _2018-10-19_24.txt | 0,56 | 0,27 | 0,45 |
| 057.ua _2018-10-19_25.txt | 0,61 | 0,25 | 0,48 |
| 057.ua _2018-10-19_26.txt | 0,55 | 0,20 | 0,39 |
| 057.ua _2018-10-19_27.txt | 0,70 | 0,28 | 0,41 |

Having analyzed the obtained results, we could conclude that the average value of cosine similarity between the texts of the trained corpus and texts of arbitrary subjects is usually within 0.35 < average ($sim_{cos}$) < 0.50. The maximum and minimum values are below 0.76 and 0.30, respectively: max ($sim_{cos}$) < 0.76 and min ($sim_{cos}$)< 0.30.

On the basis of the experimental research, we formulated the hypothesis that if the average value of the cosine similarity coefficient between the input document and documents of the trained corpus is more than 0,50 this document can be attributed to the highly specialized documents, which contain criminal information.

In order to evaluate the correctness and reliability of the obtained borderline value of the semantic similarity coefficient, we used the metrics of recall, precision and F-measure. As a result of the experiment, we analyzed 1064 documents from the test corpus, that were not used earlier, 520 of which were defined in advance as having criminally significant information, and 544 - other thematic areas.

Table 3 shows the fragment of the quality assessment table of the proposed technology to determine the semantic similarity to highly specialized texts.

**Table 3.** The fragment of quality assessment of our technology (where: *NC* – not criminal text; *C* – criminal text)

| File | Apriori information | max ($sim_{cos}$) | min ($sim_{cos}$) | Average ($sim_{cos}$) | System conclusion |
|---|---|---|---|---|---|
| 057.ua_2018-11-09_51.txt | NC | 0,65 | 0,19 | 0,42 | NC |
| 057.ua _2018-11-09_52.txt | C | 0,71 | 0,12 | 0,49 | NC |
| 057.ua_2018-10-02_17.txt | C | 0,81 | 0,22 | 0,59 | C |
| 057.ua _2018-11-09_53.txt | NC | 0,78 | 0,14 | 0,52 | NC |
| 057.ua _2018-10-02_19.txt | C | 0,82 | 0,25 | 0,67 | C |
| Misto X_2018-10-02_20.txt | C | 0,83 | 0,23 | 0,65 | C |
| Misto X_2018-11-09_54.txt | NC | 0,72 | 0,21 | 0,53 | C |
| Misto X_2018-11-09_55.txt | NC | 0,61 | 0,17 | 0,41 | NC |
| Misto X_2018-11-09_56.txt | NC | 0,79 | 0,22 | 0,43 | NC |
| Mediaport_2018-08-31_20.txt | C | 0,87 | 0,36 | 0,71 | C |
| Mediaport_2018-08-31_16.txt | C | 0,80 | 0,47 | 0,69 | C |
| Mediaport_2018-08-31_16.txt | C | 0,77 | 0,41 | 0,67 | C |
| Mediaport_2018-08-31_18.txt | C | 0,71 | 0,41 | 0,61 | C |
| Mediaport_2018-07-01_19.txt | C | 0,65 | 0,47 | 0,58 | C |
| Mediaport_2018-11-08_20.txt | C | 0,73 | 0,38 | 0,59 | C |
| ATN_2018-11-09_57.txt | NC | 0,61 | 0,15 | 0,44 | NC |

| | | | | | |
|---|---|---|---|---|---|
| ATN _2018-11-09_58.txt | C | 0,66 | 0,25 | 0,48 | NC |
| ATN_2018-11-09_59.txt | NC | 0,61 | 0,26 | 0,46 | NC |
| ATN_2018-11-09_60.txt | C | 0,79 | 0,40 | 0,62 | C |
| ATN_2018-09-04_16.txt | C | 0,88 | 0,32 | 0,70 | C |
| VKh_2018-11-09_17.txt | C | 0,85 | 0,41 | 0,72 | C |
| VKh _2018-11-09_18.txt | C | 0,78 | 0,49 | 0,69 | C |
| VKh _2018-11-09_19.txt | C | 0,84 | 0,43 | 0,71 | C |
| VKh _2018-11-09_15.txt | C | 0,87 | 0,41 | 0,71 | C |
| VKh _2018-09-04_74.txt | NC | 0,57 | 0,21 | 0,37 | NC |
| VKh _2018-11-09_62.txt | NC | 0,70 | 0,24 | 0,54 | C |
| VKh _2018-11-09_63.txt | NC | 0,51 | 0,17 | 0,39 | NC |
| VKh _2018-11-09_64.txt | NC | 0,64 | 0,23 | 0,43 | NC |
| VKh _2018-11-09_65.txt | NC | 0,53 | 0,19 | 0,37 | NC |

The result of the experiment can be presented in Table 4, where:

tp (true positive) - texts, which are correctly automatically defined as semantically close to the “criminal” theme;

fp (false positives) - texts which are incorrectly automatically defined as semantically close to the “criminal” theme;

fn (false negatives) - texts which are incorrectly automatically defined as semantically not close to the documents of “criminal” theme;

tn (true negatives) - texts which are correctly automatically defined as not close to the “criminal” theme.

**Table 4.** The results of the experiment to determine the semantic similarity of the document to the “criminal” theme

| | |
|---|---|
| tp = 512 | fp =36 |
| fn =8 | tn =518 |

Based on the above-mentioned values, we calculated the recall, precision and F-measure of the developed technology for determining the semantic similarity of the document to a highly specialized area (on the example of the criminal texts corpus).

**Table 5.** The recall, precision and F-measure of the developed technology

| *precision* | *recall* | $F_1$*-measure* |
|---|---|---|
| 93,4% | 98,5% | 95,95% |

Recall obtained from the experiments is a bit more than precision. The practical significance of these results lies in the fact that when solving this specific task of identifying criminally significant texts, it is better to have the error of the first type, which creates redundancy of criminally significant documents, than the error of the second type, which skips the criminal-contained texts.

## 6 Conclusions

The evaluation of the semantic similarity of the texts is a rather capacious and extensive task, which is an integral part of most linguistic tasks, for example, referencing, classification, creation of question-answering systems, information retrieval, etc. Most modern researches still focus on the development of this area specifically for the English language. There are a few available applications for semantic comparison of texts such as WordNet::Similarity or Alchemy API. All of them have achieved good enough results, but despite this algorithm for the other languages is still not completed. That's why the purpose of our research is to determine the thematic domain of Ukrainian and Russian texts in the absence of predefined classes by estimating semantic similarity. We consider such thematic field as a criminal-contained text and, accordingly, everything that does not fall under such topics. For the study there was created a special news corpus, all the texts of which were automatically collected from the news sites of Kharkiv in a large collection of documents (195,000 texts).

Machine learning offers many ways to define semantic similarity of texts. VSM is one of the most common and frequently used methods. But it has its disadvantages, such as the result may be irrelevant documents or lack of true relevance due to the wrong weighting of terms.

In our study, we use PPMI as a weight function to avoid this problem and to get the best results. In contrast to mutual information it refers to single events, whereas MI (mutual information) refers to the average of all possible events.

Standard measures were used to evaluate results: precision, recall and F-measure. The recall of the document classification as a highly specialized subject is above 98%, while precision is around 93% and F-measure = 96%. These values indicate good and accurate results for the application of PPMI in the task of evaluating the semantic similarity of texts.

### References


1. Salton, G., Wong, A., & Yang, C.-S.: A vector space model for automatic indexing. Communications of the ACM, 18 (11), 613–620 (1975).
2. Yoon, Kim: Convolutional neural networks for sentence classification. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1746-1751, Doha, Qatar (2014).
3. Zaidi, N. A. S. , Mustapha, A., Mostafa, S. A., Razali, M. N. : A Classification Approach for Crime Prediction. In: Applied Computing to Support Industry: Innovation and Technology, pp. 68-78. Springer, Heidelberg (2019).
4. Sayali, D. Jadhav, Channe, H.: Comparative Study of K-NN, Naive Bayes and Decision Tree Classification Techniques. In: Journal of Physics Conference Series, 1142(1):012011 (2016).
5. Rizun, N., Taranenko, Y., Waloszek, W.: Improving the accuracy in sentiment classification in the light of modelling the latent semantic relations. Information MDPI 2018. V.9. 307 (2018).

6. Bajwa, A. M., Collarana, D., Vidal, M.-E.: Interaction Network Analysis Using Semantic Similarity Based on Translation Embeddings. In: International Conference on Semantic Systems. SEMANTiCS 2019: Semantic Systems. The Power of AI and Knowledge Graphs, pp. 249-255 (2019).
7. Ehrmann, M., Talmi, J.: Starting from a blank page? Semantic similarity in central bank communication and market volatility. ECB Working Paper, No. 2023 (2017).
8. Araque, O., Zhu, G., Iglesias, C. A.: A semantic similarity-based perspective of affect lexicons for sentiment analysis. In: Knowledge-Based Systems, Volume 165, pp. 346-359 (2019).
9. Ming, Liu, Bo, Lang, Zepeng, Gu: Calculating Semantic Similarity between Academic Articles using Topic Event and Ontology. Published in ArXiv (2017).
10. Ramage, D., Rafferty, A. N., Manning, C. D.: Random walks for text semantic similarity. In: Proceedings of the 2009 workshop on graph-based methods for natural language processing, Association for Computational Linguistics, pp. 23–31 (2009).
11. Resnik, P.: Using information content to evaluate semantic similarity in a taxonomy. In: Proceedings of the 14th International Join Conference on Artificial Intelligence, 1995, pp. 448-453, San Mateo, CA (1995).
12. Gad, W.K, Kamel, M.S.: New semantic similarity based model for text clustering using extended gloss overlaps. In International Workshop on Machine Learning and Data Mining in Pattern Recognition, V.7., 23, pp. 663-677 (2009).
13. Turnay, P.D., Pantel, P.: From frequency to meaning: Vector Space Models of Semantics. Journal of Artificial Intelligence Research, 37, pp. 141-188 (2010).
14. Majumder, G., Pakray, P., Gelbukh, A., Pinto, D.: Semantic Textual Similarity Methods, Tools, and Applications: A Survey. Comp. Sist. vol.20, no.4, México (2016).
15. Furnas, G.W., Landauer, T.K., Gomez, L.M., & Dumais, S.T.: Statistical semantics: Analysis of the potential performance of keyword information systems. Bell System Technical Journal, 62 (6), pp. 1753–1806 (1983).
16. Pantel, P., Lin, D.: Document clustering with committees. In: Proceedings of the 25th Annual International ACM SIGIR Conference, pp. 199–206 (2002).
17. Sidorov, G., Gelbukh, A., Gómez-Adorno, H., Pinto, D.: Soft similarity and soft cosine measure: Similarity of features in vector space mode. Computación y Sistemas. V.18, 3, pp. 491-504 (2014).
18. Khairova, N., Kolesnyk, A., Mamyrbayev, O., Mukhsina, K.: The aligned Kazakh-Russian parallel corpus focused on the criminal theme. In: Proceedings Colins 2019, pp. 116-125, Ukraine, Kharkiv (2019).