UDC 332.7:[004+316.472.4]

Sinyak N.G.
Tajinder S.
Madhu K.J.
Kozlovskiy V.V.

Predicting real estate market trends and value using pre-processing and sentiment text mining analysis

Ubiquitous growth in the text mining field is unprecedented, where social media mining is playing a significant role. Gigantic growth of text mining is becoming a potential source of crowd wisdom extraction and analysis especially in terms of text pre-processing and sentiment analysis. The analysis of a potential influence of sentiment on real estate markets controversially discussed by scholars of finance, valuation and market efficiency supporters. Therefore, it's a significant task of current research purview which not only provide an appropriate platform for the contributors but also for active real estate market information seekers. Text mining has gained the widespread attention of real estate market information users which is almost on explosion level. Accessibility of data on such behemoth scale mandates regular and critical analysis of this information for various perspectives' plausibility. Rich patterns of online social text can be exploited to extract the relevant real estate information effectively. As text mining plays a significant and crucial role in discovery of these insights therefore its challenges and contribution in social media analysis must be explored extensively. In this paper, we provide a brief about the current summary of the modern state of text mining in pre-processing and sentiment for the real estate market analysis. Emphasis is placed on the resources and learning mechanism available to real estate researchers and practitioners, as well as the major text mining tasks of interest to the community. Thus, the main aim of this chapter is to expound and intellectualize the domains of social media which are accessible on an extraordinary range in the field of text mining real estate for predicting real estate market trends and value.

Keywords: *housing, real estate, sentiment analysis, social media, text mining, spam, CRF, n-gram, entropy*

Sinyak Nikolay Georgievich,
PhD in Economics, Professor of the Department of Management and Economics, Private Institute of Management and Business; bldg. 3, 1 Slavinsky St., Minsk, 220086, Belarus;
SPIN-code: 4400-2224,
Scopus: 55952470200,
ResearcherID: K-4838-2015;
ORCID: 0000-0002-1688-9268,
Google Scholar: 9wEDUrMAAAAJ;
siniakn@gmail.com

Tajinder Singh,
Assistant Professor of Computer Science and Engineering Department; Sant Longowal Institute of Engineering and Technology; SLIET Administration Block, SLIET Rd, Sangrur, Punjab, 148106, India;
Google Scholar: nn4t8UMAAAAJ&hl;
nith2k14@gmail.com

Madhu Kumari Jaglan,
Assistant Professor of Computer Science & Engineering Department, National Institute of Technology Hamirpur; Himachal Pradesh, Hamirpur, 177005, India;
ORCID: 0000-0003-3203-2579;
Madhu.jaglan@gmail.com

Kozlovskiy Vitaliy Vladimirovich,
Professor, Chief Researcher; Research Economic Institute of the Ministry of Economy of the Republic of Belarus; bldg. 1, 1 Slavinsky St., Minsk, 220086, Belarus; SPIN-code: 4354-7264,
ResearcherID: ABI-8122-2020,
ORCID: 000-0002-9194-6170;
vital_kozlovsky@mail.ru

1. INTRODUCTION

Nowadays, the Web is unique a foremost assets for text corpora because of the enormous extent of HTML documentations containing text of all manners. Thus, text mining is gaining a great attention in recent years due to huge amount of text which is posted in a number of social media, networks, blogs and forums. It is a knowledge intensive manner in which user of social media deals with text collected over time with the help of analysis tools. As data mining, similarly text mining seeks to extract suitable information from text sources through the identification and exploration of fascinating patterns. However, the text sources are text collections in text mining and useful, valuable patterns are found not among formalized database records but in the unstructured textual data whereas, data mining assumes that data have already been stored in a structured format, much of its preprocessing focus falls on two critical tasks: scrubbing and normalizing data and creating extensive numbers of table joins. Online social media has generated innovative pattern of information sharing which not solitary offers platform for the contributors but also for dynamic, active information seekers. Numerous types of social media have gained the extensive attention internet users' almost on explosion level. Availability of data on such behemoth scale mandates regular and critical analysis of this information for various perspectives' plausibility. In past few years, number of mechanism and training procedures is presented which are used to share and communicate with other users in the form of text only and such systems also handles the recurrence of text [1]. Such mechanisms are providing a conducive milieu for all social users to generate and diffuse information with which the intensification of usage of social media is rapidly growing.

With the unexpected evolution of information in the World Wide Web, with the growth of online user's, online text increasingly turns into the foremost source for creating eminence corpus of huge size. From WWW, Leipzig Corpora Collection [2] helps to combines various schemes for assembling textual data. This collected text is used for preprocessing [3] from which a valuable, knowledgeable, feature based textual information is extracted which are used for further processing such as sentiment analysis [4, 5], sentiment polarity disambiguation [6], event detection [7, 8] and classification [9], trend prediction [10], topic tracking [11], Recommendation system [12], etc. Furthermore, various levels of text representation are also illustrated in which text can be simply be represent as bag-of-words, string of words or it can be signified semantically so that further expressive, evocative analysis and mining can be done. Name entity recognition is a good example of information extraction [13]. Thus, the main aim behind text mining is to convert large corpus of text into numbers by applying influential mining technique to extract meaningful knowledge patterns [14]. With this to bring clarity in the field of text mining, seven foremost areas within text mining are playing significant role which are:

1. *Search and information retrieval (IR):* it helps to search and retrieve documents from huge corpus according to keyword queries, covers indexing [14, 15].

2. *Document clustering:* it helps to collect related documents into clusters in text mining with the help of algorithm [15–17].

3. *Document classification:* it is the supreme prominent method used in text mining which uses a model of text learned from documents with well-known labels to untagged documents [15, 18].

4. *Web mining:* due to distinctive structure and huge availability of text on web typically in the form of tags, hyperlinks within document (usually in structured manner), used to extract knowledgeable patterns i.e. web content, web structure, and web usage data [15, 19].

5. *Information extraction (IE):* structured text is to be created from collected unstructured or semi structured text [13, 15].

6. *Natural language processing (NLP):* it is an influential tool and in text mining it provides valuable information such as part of speech tags, phrase boundaries etc. [15, 20].

7. *Concept extraction:* it is to detecting relationship between texts in the text based on a lexical database. It matches the words which generate closely matched conceptual [15, 21].

Furthermore, such deep-learning classifiers as artificial neural networks have the prospective to extract a much richer information structure from textual documents. They provide more data available for training with a better scalability and are predetermined for real time analytics and big data applications, which further estimates them superior to traditional indicators.

In spite of these advantages, the potential of textual sentiment indicators in real estate has not been enough theoretically explored. With respect to traditional data and value analysis in real estate, textual sentiment, machine- and-deep-learning classifiers have been absolutely ignored.

1.1. Activities of text mining process (Fig. 1)

Activity 1. Establish the Corpus. In text mining quality and quantity of text are supreme elements. The main aim of this activity is to bring together all the documents that are appropriate to the research problem being addressed. Occasionally, collected documents in the text mining are readily obtainable which can be used by the research depiction (e.g., conducting sentiment analysis on customer reviews of a precise product/service).

Activity 2. Preprocess the Data. This phase helps to generate structured representation of collected corpus of text. It plays an important role in text mining as its aim is to tame text in terms of noise [1, 22] is emoticons, misspelled or incorrect.

Activity 3. Extract the Knowledge. Innovative, knowledgeable, patterns are take out from the processed text in the milieu of precise problem being addressed using the structured text. Predictions (e.g., classification, regression, and time-series exploration), Clustering (e.g., segmentation and outlier analysis), Association (e.g., affinity analysis, link analysis, and sequence analysis), trend analysis are the core classifications of information extraction techniques in text mining.

2. MOTIVATION

As text mining is becoming popular and advanced in current era thus, in this way it has become integral part of human lives. Text in the social media is usually huge in size, with multi-modality capabilities which allows them to collect actual feature of text from raw text. The features are deployed according to the need or thousand leading towards efficient monitoring at higher granularity as compared to traditional text mining. Due to intelligent and cooperating behavior of pre-processing task it is easier to perform complex task in unlabeled, raw and from dynamic text efficiently. Usually in text mining contextual information is not clear as ambiguities exist between textual words which lead users to take wrong decision. Thus in text mining to capture domain knowledge, learning techniques should be efficient which cope up with missing, noisy and dynamic nature of textual data. In this way it leads to more accurate

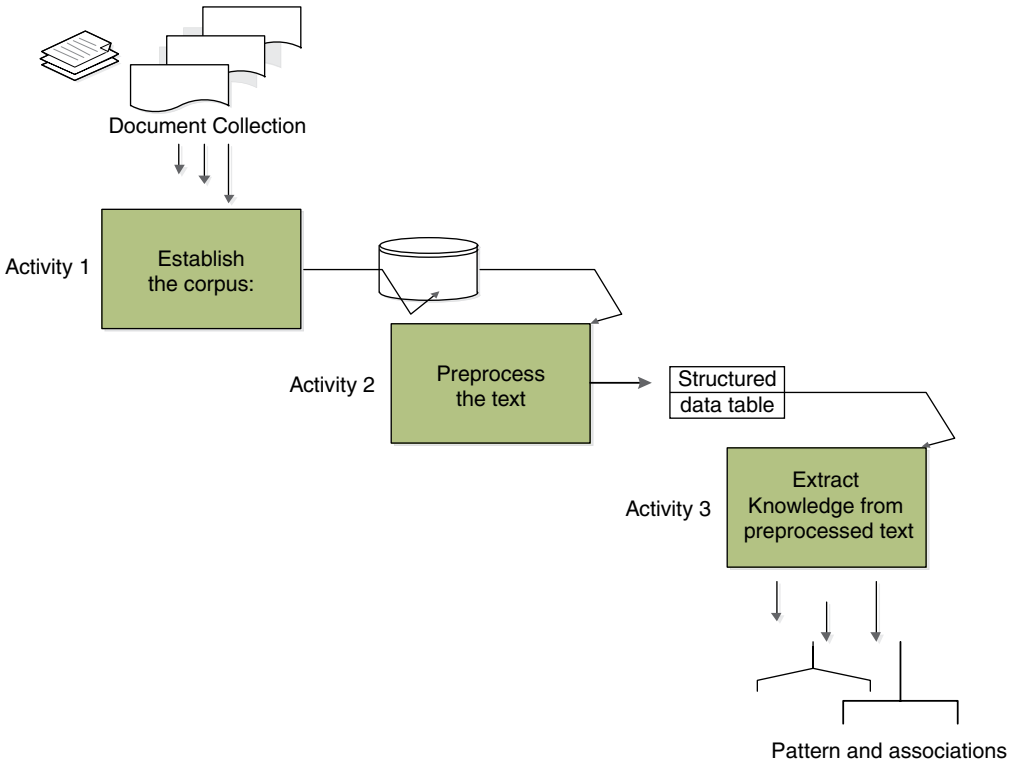

Fig. 1. A detailed view of the context diagram for text mining

tools for extracting semantic as well as contextual information which provides empirical properties of social text and social interactions of users. With advancements of text mining, the applications are no longer limited to just corpus monitoring rather extends towards event detection, classification, trend prediction and estimation using text streams also. Another certain innovative and exciting applications in text mining are also present such as Semantic Mapping and Web search enhancement, which promise have progressively essential effects of the depth of analysis and quality of text mining results.

2.1. Applications of text mining

Text mining is used for a wide array of applications spanning multiple domains. Broadly applications can be classified according to the goals that are tough to define in general terms. Compared to other approaches and established methods, text mining is advance and unstandardized analytical technique for knowledge discovery. Therefore variety of methods, procedures, techniques and tools are used to extract knowledge from textual data. Text mining is well established in Semantic Mapping, sentiment analysis and opinion mining, Mining Bibliographic Data, Enhancing Web Search, trend detection, event detection and classification, recommendation system and many more. These ever increasing applications of text mining can be explained in detailed way as given below:

Open ended survey. In marketing it's an integral part of mining and analyzing the customer's attitude and behavior using open ended questions refers to the theme under exploration. Main aim is to permit respondents to express their interpretations or sentiments without obliging them to specific format or extents.

Involuntary processing of posts, messages, emails, etc. Automatic classification of text is a common application of text mining in which most undesirable "junk email" is possible to filter out. Similarly, messages, posts can automatically be discarded.

Application in evaluating warranty, insurance entitlements and analytical interviews. In various business domains vast amount of information is collected using open ended textual form. Thus in estimating warranties entitlement and medical insurance, brief interviews can be summarized which is collected electronically. Then such kind of narratives is read by text mining approaches for further processing and in this way they play a vast role in text mining applications.

Exploring challengers/competitors by crawling web sites. In a specific domain, automatically processing of web contents of web pages is an emerging application. It helps to derive list of terms and documents available at sites and supports to determine useful, knowledgeable features that are described.

Semantic mapping. Dimensions between word illustrations allow the mapping of indirect relationships between words that appear in related contexts. Semantic mapping is not a new concept in text mining. Latent semantic indexing (LSI) is playing significance role to correlate semantically related terms that are latent in a collection of text documents and latent semantic analysis (LSA) feats the co-occurrence of words in text segments respectively.

Enhancing web explorations. In modern development, meta search engines are gaining great momentum in web searches. These search engines enhance the web exploration where one search engine sends out search request to other search engine and displays all of the return. It offers an opportunity to browse web specifically to enhance the consequences of browsing session.

Sentiment classification. The arena of sentiment analysis deals with the analysis of opinions found in text documents. In this case text is sorted into positive, negative and neutral classes. Polarity disambiguation at sentence and document level is an emerging area of research in which contextual polarity at document and as well as sentence level is to be analyzed to find the actual score of sentiment. Various shopping and survey sites use the scale rating in the form of (★★★★★) or (1 to 10) to collect the customer's sentiments but sentiments in the form of text conveys actual experience of the customers.

Mining Bibliographic Data. It is based on to extract text, related to bibliographic and distinguishes duplicate references. It helps to arrange and establish co-authors relationships and can further be used for analysis by visualizing the functionalities of extracted bibliographic text.

Text mining in detecting Antagonistic, Heated Language in Mails and Cybercrime. In text mining few researchers are dealing with the design of precise application which detects antagonistic from group chat or blog. Use of a bag of words and a statistical classification approaches are playing a key role in text mining to deal with heated languages and to collect chat text from the chat rooms to prevent the cybercrime.

Detecting Popular Events and Trend Prediction. Social media platform offers abundant chance to post on social media, to participate in social forums, events and blogs. Thus its great opportunity for researchers to analyze the event on social media as well as to find the impact of event on trend can also be evaluated. Novelty of event detection and classification is a hot application in text mining which captures the newly arrived with other people on single platform about famous events and trends.

Profile Spams Detection (Fig. 2). In text mining spams profile detection is an emerging application where a blacklist of URLs, DNS, IP address etc. is to be generated and if fake profile match with this database then it will be blocked automatically. Classification approaches are used to classify the spam profiles, links, URLs etc. which blocks the spammers to distribute spam messages, promoting personal blogs, advertisements, pages etc. Fig. 2. is demonstrating the framework of spam detection.

2.2. Design issues in text mining

In text mining majority of text is in unstructured form and it is most common type of text which appears in emails, web pages, blogs, etc. thus this unstructured data demands to be processed for further decision making and then pre-processing come into real picture. Generally a computer cannot understand about the text being communicated, simply structure and syntax of the text is defined. Syntax refers to the organization of language and how distinct words are collected to create well rounded sentences. Whereas semantic define the importance of specific words within the neighboring context. Thus, in text mining it's important to understand the contextual information to analyze the actual meaning of various keywords in the given sentence and document. Following are few of the important issues that must be taken care while designing and studying the text mining:

2.2.1. *Multi-Modality*

Usually, text in the social media contains information which is not required for processing. As it may contains text, images, tags, URLs and hyperlinks. Most of the text mining approaches deal with text only. Thus, it is essential to eliminate undesirable additional information from the collected data [3] otherwise it leads to social text noisy and ambiguous. This is a big challenge in text mining to access

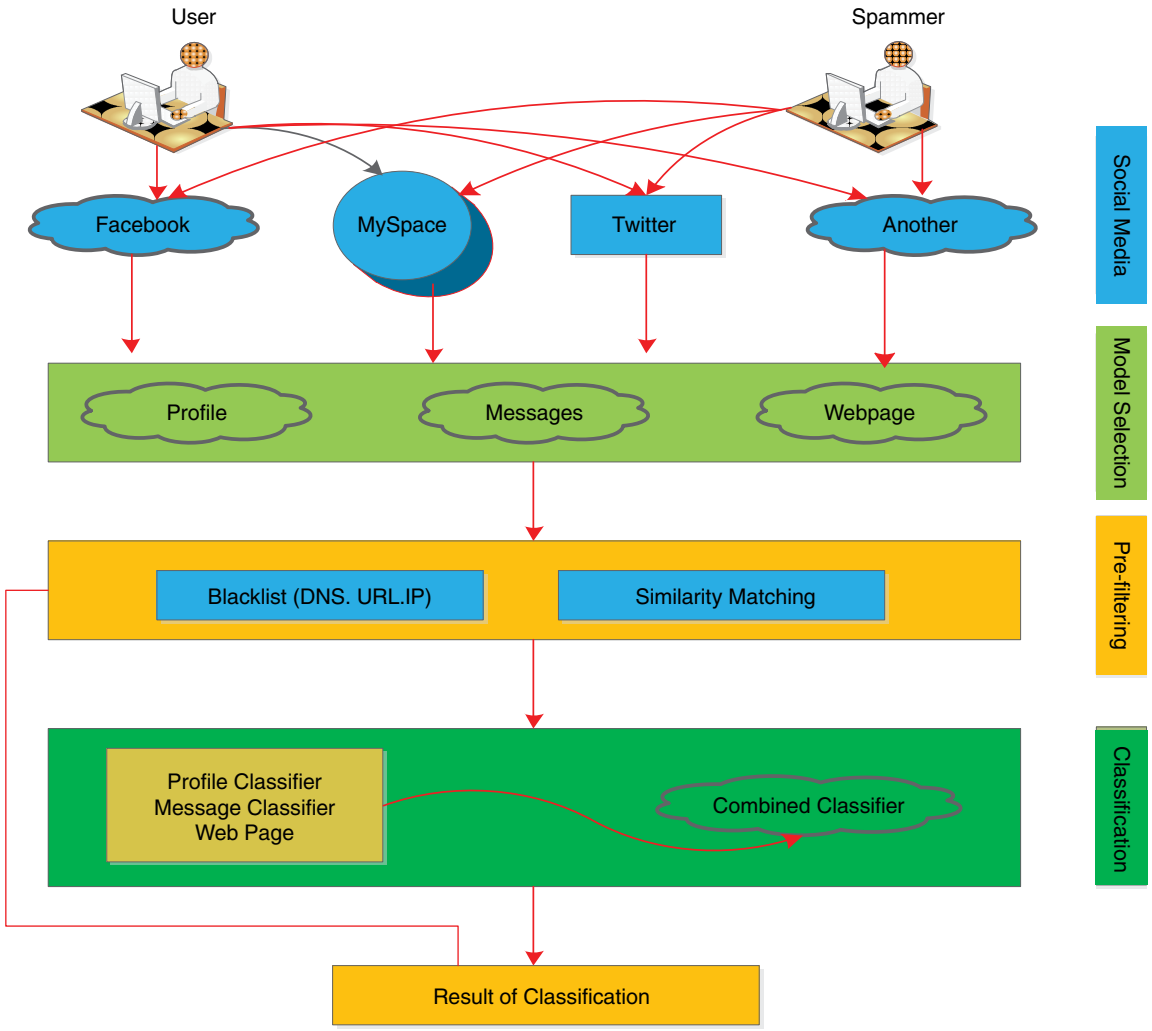

Fig. 2. Framework for Spam Detection Source

useful information buried in collected corpus of text which is stored in different forms (text + image + tags + hyperlinks). Information extraction and information retrieval are used to access information from a sequence of documents whose aim is to collect information which can be used further for sentiment analysis [4, 23], event detection [7] and classification, trend measurement [10], etc.

2.2.2. *Rich social context*

Text mining is rich of social context which is also known as social environment where the online social users interact with each other's in groups, chat rooms, blogs etc. This rich nature of social text generates number of textual connection, graphs and tree structures from root word to the stem word. From the social graph and tree structure of text features of text can be obtained but it is a major issue in text mining to classify text on the basis of features and context. Thus number of parsing mechanism (dependency parsing) is used to find the actual structure of collected text.

2.2.3. *Inconsistent quality*

Usually, user produced information on social media is categorized by noise because users use number of slangs, abbreviations, emoticons, short form of words etc. which acts as a noise in text mining. Social media is occupied with spam messages, frauds, as well as internet viruses. Thus textual information in social media is very limited and the available textual information usually noisy which needs to be processed before further processing. Thus needs a text mining pre-processing schemes to handle such noisy textual information in social media.

2.2.4. *Huge volume and multi-source*

Text in the social media may be generated by various applications, blogs and users' communities which is produced tremendously and spread over social media. It may take less time to diffuse over media or may take long time which depends on the active participation of users. Collected text will be in un-structured form and it can be in the form of corpus or from stream with huge volume based on predefined query or keyword. Thus it's a big issue in text mining to handle text generated from numerous sources which is huge in capacity.

2.2.5. *Dynamic nature of social text*

Due to dynamic nature of social media in real time scenario users post in very less span of time which needs online fast clustering to target fast updates and to cluster together all upcoming text in less amount of time and accurately. Thus, to maintain quality/correctness of collected data and to prevent resources from wastage because of fast updates, the text should be identified and handled timely in the real time situation.

2.2.6. *Slangs and folksonomies*

Generally, in text mining different people have dissimilar symbolization/notations for information sharing. Number of slangs and folksonomies are used in text mining which is corrected by WordNet dictionary for matching process to get the actual meaning of the

particular word and in which context it is represented. Therefore, in order to sustain the correctness of the words and to provide desired meaning within the text, the timely identification of word sense and their recovery is necessary. Due to isolated conditions, the text process should have capabilities to cope up with such correctness without any human intervention.

3. GENERAL OUTLINE OF TEXT PRE-PROCESSING

Text pre-processing (Fig. 3) contains number of steps in a sequence designed at converting noisy text into an appropriate form for input to an algorithm. Fig. 3 illustrate the distinctive operations of text pre-processing for sentiment analysis. Tokenization is most important step in sentiment analysis because it represents sentiment information sparsely and unusually. In tokenization the input text is divided into its small units and the subsequent tokens are tagged with their respective POS and sentiment scores are extracted from SentiWordNet.in the next phase, tokens are transformed to a reliable case using lemmatization and finally, filtering of stop word is typically implemented.

3.1. Main Challenges related to Social Media Text

Social media is ambiguous notion which refers to websites and applications that facilitate consumers

to generate and share text or contribute in social networking [24]. In modern era, social media also refers to social networking sites such as Facebook, G+, and LinkedIn [25]. Website and applications of social media comprise delicate blogs, micro-blogs, web forums, community question-answering, mailing lists, and numerous websites through social networking facilities [25, 26]. Broadly social media is defined which include extensively accessible electronic tools that allow everyone to distribute, access, and broadcast information. Social networking is an essential feature of social media. In [27], authors divide social media into eight types such as: blogs, micro-blogs, e-commerce portals, multimedia sharing, social networks, review platforms, social gaming, and virtual worlds. Documents of social media contain exclusive features in numerous aspects such as:

1. **Shortness.** With the comparison of traditional media, social media documents usually shorter. Consider an example of twitter then, there is a 140-character limit to the length of a tweet [28]. Thus, it's very challenging to extract important features from the short length of tweet.

2. **Multi-linguality.** In social media, users use different languages to participate in various communities, blogs, and groups in the identical communication platform. Due to this multi-linguality nature of social media, usually it becomes difficult to analyze the text message as it will be out of vocabulary. E.g. in FIFA WorldCup 2014, people deliberate the identical match in several languages on Twitter [25].

3. **Opinions.** Numerous documents in social media hold opinions. An opinion is a quintuple (o_j, f_{jk}, so_{ijkl}, h_i, t_l), where, o_j is a target object, f_{jk} is a feature of the object o_j, so_{ijkl} is the sentiment value of the opinion of the opinion holder h_i on feature f_{jk} of object o_j at time t_l. so_{ijkl} is positive, negative or neutral, or a more granular score, h_i is an opinion holder, t_l is the time when the opinion is expressed [29].

4. **Timeliness**. Social media is dynamic in nature and documents are posted with precise timestamps. Thus, the text over social media changes over time in social text streams. Event detection, classification and trend prediction in text stream is quite challenging.

Table 1. Social text quality challenges

Challenge	Description
Stop list	Common words frequency of occurrence
Lemmatization	Similarity detection of text/words
Text cleaning	Removal of unwanted from the data
Clarity of words	To clear the meaning in text
Tagging	Predicting data annotation and its characteristics
Syntax /Grammar	Scope of ambiguity, data dependency
Tokenization	Various methods to tokenize words or phrases
Automatic learning	Similarity measures and use of characterization

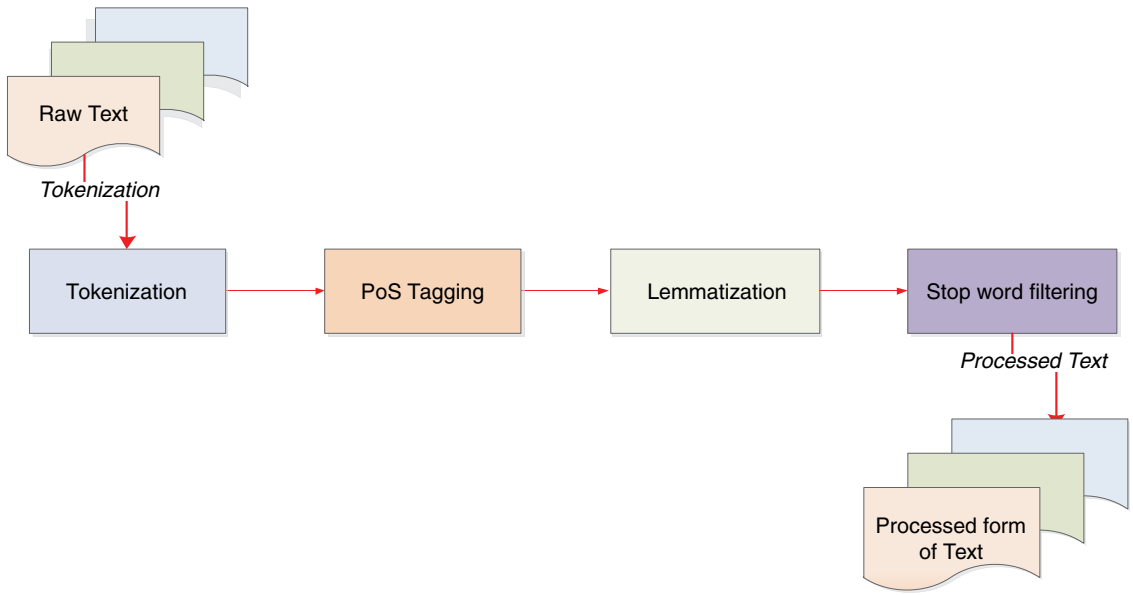

Fig. 3. General outline of text pre-processing

4. TEXT PRE-PROCESSING

In social media, non-standard words show their negative impact, thus normalization methods are persistently emerging. For instance, an ISBN is generally used to represent "*International Standard Book Number*" [30]. Mobile phone SMS messages often hold non-standard words like *hru*/how are you, *ttyl*/talk to you later, *cu*\see you [26] and many more. Thus with the growth of social media, twitter contains a huge non-standard words, including *mistakes/typos* (e.g. *helo*/hello", *stndrd*/standard"), phonetic approximations (e.g., *w8*\wait", *f9*/fine"), words with repetitions (e.g. *hiiiiiiii*/hi", *helloooooooo*/hello"), informal abbreviations (e.g., *awsummmmmm*/awesome", *gr888888888*/great"). Such non-standard words act as noise in social media text mining and it is hard and challenging task to compute all possible factors underlying the formation of non-standard words.

4.1. Analysis of text pre-processing (The n-gram model and assumptions)

The *n*-gram model is a language model that allocates probabilities to series of words. It is a sequence of n words, whereas a 3-gram (trigram) contains three-word sequence like "with humble submission", or "humble submission and" and a 2-gram (bigram) contains two-word sequence like "with humble" or "humble submission". Now, we will understand the use of *n*-gram model to evaluate the probability of the preceding word of an *N*-gram specified the prior words, and furthermore to allocate probabilities to whole word sequences. In estimating the probabilities of the next words or the while word sequence, *n*-gram model is the greatest model in language processing. With *n*-gram probability of entire sequence like $P(t_1, t_2, \ldots, t_n)$ is computed as:

$$P(X_1 \ldots X_n) = P(X_1)P(X_2|X_1)P(X_3|X_1^2)\ldots P(X_n|X_1^{n-1}) = \prod_{k=1}^{n} P(X_k|X_1^{k-1}). \quad (1)$$

Thus, after applying chain rule it will be computed as:

$$P(t_1^n) = P(t_1)P(t_2|t_1)P(t_3|t_1^2)\ldots P(t_n|t_1^{n-1}) = \prod_{k=1}^{n} P(t_k|t_1^{k-1}). \quad (2)$$

The above equation (1) shows the connection between joint probability of a word sequence and calculating the conditional probability of a word specified prior words whereas, from the equation (2), joint probability of complete word sequence can be estimated by multiplying together a number of conditional probabilities.

On the other hand, if we consider 2-gram (bigram) language model and want to estimates the probability of a word given all the previous words $P(t_n|t_1^{n-1})$, with the help of conditional probability of the previous words $P(t_n|t_{n-1})$. With the use of 2-gram (bigram), conditional probability of the next word is calculated by using the following calculation:

$$P(t_n|t_1^{n-1}) \approx P(t_n|t_{n-1}). \quad (3)$$

Markov assumption. It states that the probability of a word based solely on the previous word is known as Markov assumption. It is supposed that, the probability of future words can be presume without recognizing too deep into the past. Thus, we can simplify the 2-gram, which inspect one word into the prior, whereas the 3-gram, inspect two word into the prior and as a result to the *n*-gram, which inspect $n - 1$ word into the prior. Therefore, the probability of the consequent word in a word sequence in *n*-gram to the conditional probability is described as:

$$P(t_n|t_1^{n-1}) \approx P(t_n|t_{n-N+1}^{n-1}). \quad (4)$$

In case of 2-gram, from probability of a specific word probability of a whole word sequence can be computed by replacing eq. (3) into eq. (2) as:

$$P(t_1^n) \approx \prod_{k=1}^{n} P(t_k|t_{k-1}). \quad (5)$$

4.2. Conditional random fields (CRF) and assumptions

The philosophy of conditional random fields is too profound to be explained. These are used in number of applications where to calculate several variables that be influenced on each other. Conditional random fields (CRFs) are a structured forecast framework which is widely used for name entity recognition (NER). These are the probalistic framework, which are used for labeling and segmenting organized well-structured text, such as classification, sequences and trees. The fundamental task is that of describing a conditional probability dispensation over train text given a particular consideration train, comparatively a combined distribution over both train and consideration train. Moreover, CRFs are undirected graphical models much flexible than hidden Markov models as they abstain label bias issue. They perform better in both MEMMs and HMMs including many real world efforts in various fields such as bioinformatics, linguistics etc. in text classification and sequence labeling, CRFs provides much flexibility to define the feature vector.

Definition. Consider $G = (V, E)$ be a graph so that $Y = (Y_v)_{v \in V}$, thus, Y is ordered by the vertices of G. Therefore, (X, Y) are the CRFs in case, when trained on X, the random variables Y_v follow the estate of Markov w.r.t. the graph which is given below (eq. 8) in which, $w \sim v$ indicates that in G, w and v are neighbor:

$$p(Y_v|X, Y_w, w \neq v) = p(Y_v|X, Y_w, w \sim v). \quad (6)$$

According to the essential proposition of random arenas, joint distribution over the label order sequence Y given X is described as eq. 9, in which, x and y are the data sequence and label sequence respectively:

$$p_\theta(y|x) \propto \exp\left(\sum_{e \in E,k} \lambda_k f_k(e, y|_e, x) + \sum_{v \in V,k} \mu_k g_k(v, y|_v, x)\right). \quad (7)$$

Moreover, in defining the features function, a set of real value feature is created by which some features and characteristics of the training data set are described which is defined as (eq. (8)):

$$W(x,i) = \begin{cases} 1 & \textit{when the observation is at the position } i; \\ 0 & \textit{otherwise.} \end{cases} \quad (8)$$

Fig. 4 defines the depiction of functional model called conditional random field and the eq. (9) elaborate the CRFs as a functional unit:

$$p(y|x) = \frac{\exp\left\{\sum_{m=1}^{M} \lambda_m f_m(y_n, y_{n-1}, x_n)\right\}}{\sum_{y'} \exp\left\{\sum_{m=1}^{M} \lambda_m f_m(y'_n, x_n)\right\}}. \quad (9)$$

Thus, in general, a linear condition random field (CRF) is defined as a distribution of $p(Y|X)$ that takes the form as in eq. (10), in which $Z(X)$ is known as instance specific normalization function:

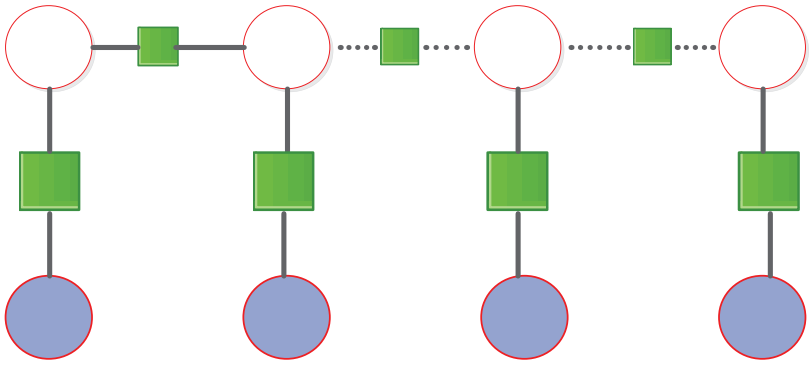

Fig. 4. Functional model of conditional random fields (CRFs)

$$p(Y|X) = \frac{1}{Z(X)} \exp\left\{\sum_{m=1}^{M} \lambda_m f_m (y_n, y_{n-1}, x_n)\right\}; \quad (10)$$

$$Z(X) = \sum_{y} \exp\left\{\sum_{m=1}^{M} \lambda_m f_m (y_n, y_{n-1}, x_n)\right\}. \quad (11)$$

Furthermore, conditional random fields (CRFs) have many applications and they are used to a number of fields. In text mining, CRFs are used for named entity recognition (NER). These are commonly accepted by many researches because they represent entities graphically and their dependencies are also characterized including their rich features. In this dissertation, we used conditional random fields (CRFs) for named entity recognition in tweets for the purposed of text normalization.

5. CONCLUSIONS

This article is an endeavor to elucidate text analytics bent of social media with current updates and text mining method. As social media has progressed in exceptional mode, it led to numerous motivating and exciting research direction particularly in the arena of text mining. Looking at the interestingness of the text mining area, it is full of information. Thus, the main impact of this paper is to expand and conceptualize the areas of social media which are available and can be accessible on an astonished variety. Social media text is full of noise, abbreviations, special symbols, emoticons out of vocabulary words and folksonomy. Therefore, rich patterns of text stream or corpus, can be exploited to generate relevant and required information. Number of techniques are present for text mining which can't be ignored and helps to measure the impact on text pre-processing and sentiment analysis. This article explains the various tasks along with real estate market application areas. It also discusses the various methodologies available for text pre-processing as well as for the sentiment analysis in text corpus and text stream. As a discovery, these techniques can be used for real estate market analysis and valuation and text mining can be applied for prolific research domain such as: event detection and classification, sentiment polarity disambiguation, trend prediction, semantic hashing, crowd sourcing etc. Yet, it is to be evolve few multilingual model which are capable to handle multilingual eccentricities in text stream and text corpus to facilitate the better classification in text mining. For the first time, the capabilities of a machine- and a deep-learning classifier for predicting real estate market trends are assessed. Real estate market updates and news analytics by means of textual sentiment classifiers in general and machine- and deep-learning algorithms in particular can be perceived as a valuable and innovative source of market sentiment and is able to provide real estate researchers and valuers with a reliable leading market indicators.

REFERENCES

1. Lifna C.S., Vijayalakshmi M. Identifying concept-drift in twitter streams. Procedia Computer Science. 2015; 45:86-94. DOI: 10.1016/j.procs.2015.03.093
2. Goldhahn D., Eckart T., Quasthoff U. Building large monolingual dictionaries at the Leipzig Corpora collection : From 100 to 200 languages. Proceedings of the Eighth International Conference on Language Resources and Evaluation. 2012; 759-765. URL: http://www.lrec-conf.org/proceedings/lrec2012/pdf/327_Paper.pdf
3. Singh T., Kumari M. Role of text pre-processing in twitter sentiment analysis. Procedia Computer Science. 2016; 89:549-554. DOI: 10.1016/j.procs.2016.06.095
4. Bhadane C., Dalal H., Doshi H. Sentiment analysis: Measuring opinions. Procedia Computer Science. 2015; 45:808-814. DOI: 10.1016/j.procs.2015.03.159
5. Hamdan H., Bellot P., Bechet F. Lsislif: Feature extraction and label weighting for sentiment analysis in Twitter // Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015). Association for Computational Linguistics, 2015; 568-573. DOI: 10.18653/v1/S15-2095
6. Saleiro P., Rodrigues E.M., Soares C. FEUP at SemEval-2017 Task 5 : Predicting Sentiment Polarity and Intensity with Financial Word Embeddings. 2017; 904-908.
7. Vavliakis K.N., Symeonidis A.L., Mitkas P.A. Event identification in web social media through named entity recognition and topic modeling. Data & Knowledge Engineering. 2013; 88:1-24. DOI: 10.1016/j.datak.2013.08.006
8. Zhou X., Chen L. Event detection over twitter social media streams. The VLDB Journal. 2014; 23(3):381-400. DOI: 10.1007/s00778-013-0320-3
9. Zhao Q., Mitra P. Event detection and visualization for social text streams. ICWSM 2007 — International Conference on Weblogs and Social Media. 2007; 26-28.
10. Aiello L.M., Petkos G., Martin C., Corney D., Papadopoulos S., Skraba R. et al. Sensing trending topics in twitter. IEEE Transactions on Multimedia. Institute of Electrical and Electronics Engineers, 2013; 15(6):1268-1282. DOI: 10.1109/TMM.2013.2265080
11. Lin J., Snow R., Morgan W. Smoothing techniques for adaptive online language models: Topic tracking in tweet streams. Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining — KDD '11. 2011; 422-429. DOI: 10.1145/2020408.2020476
12. Martinez L., Barranco M.J., Pérez L.P., Espinilla M. A knowledge based recommender system with multigranular linguistic information. International Journal of Computational Intelligence Systems. 2008; 1(3):225-236. DOI: 10.1080/18756891.2008.9727620
13. Quan C., Ren F. Unsupervised product feature extraction for feature-oriented opinion determination. Information Sciences. 2014; 272:16-28. DOI: 10.1016/j.ins.2014.02.063
14. Croft W.B., Metzler D., Strohman T. Search engines: Information retrieval in practice. Pearson Education, 2015; 542.
15. Miner G., Delen D., Elder J., Fast A., Hill T., Nisbet R. The seven practice areas of text analytics. Elsevier, 2012; 29-41. URL: https://www.elderresearch.com/hubfs/Whitepaper_The_Seven_Practice_Areas_of_Text_Analytics_Chapter_2_Excerpt.pdf

16. Ghai D., Gera D., Jain N. A new approach to extract text from images based on DWT and K-means clustering. International Journal of Computational Intelligence Systems. 2016; 9(5):900-916. DOI: 10.1080/18756891.2016.1237189
17. Aggarwal C.C. Chapter 10. A survey of stream clustering algorithms. Data Clusterin. 2013; 229-252. DOI: 10.1201/9781315373515-10
18. Li X., Yu W. Fast support vector machine classification for large data sets. International Journal of Computational Intelligence Systems 2013; 7(2):197-212. DOI: 10.1080/18756891.2013.868148
19. Srivastava J., Desikan P., Kumar V. Web mining — concepts, applications and research. 2002; 51-71. URL: https://www-users.cs.umn.edu/~desi0016/publications/wmo.pdf
20. Popowich F. Using text mining and natural language processing for health care claims processing. ACM SIGKDD Explorations Newsletter. 2005; 7(1):59-66. DOI: 10.1145/1089815.1089824
21. Gelfand B., Wulfekuhler M., Punch W. Automated concept extraction from plain text. AAAI Technical Report WS-98-05. 1998. URL: https://www.aaai.org/Papers/Workshops/1998/WS-98-05/WS98-05-003.pdf
22. Kumar J.P., Govindarajulu P. Near-duplicate web page detection: An efficient approach using clustering, sentence feature and fingerprinting. International Journal of Computational Intelligence Systems. 2013; 6(1):1-13. DOI: 10.1080/18756891.2013.752657
23. Bhuta S., Doshi U. A review of techniques for sentiment analysis of twitter data. 2014 International Conference on Issues and Challenges in Intelligent Computing Techniques (ICICT). IEEE, 2014; 583-591. DOI: 10.1109/ICICICT.2014.6781346
24. Peetz M.-H. Time-aware online reputation analysis. 2015. URL: https://pure.uva.nl/ws/files/2550354/157890_peetz_thesis_complete.pdf
25. Ren Z. Monitoring social media: Summarization, classification and recommendation. 2016. URL: https://hdl.handle.net/11245/1.541961
26. Han B. Improving the utility of social media with natural language processing : PhD thesis. 2014. URL: http://hdl.handle.net/11343/41029
27. Aichner T., Jacob F. Measuring the degree of corporate social media use. International Journal of Market Research. 2015; 57(2):257-276. DOI: 10.2501/IJMR-2015-018
28. Ren Z., Liang S., Meij E., de Rijke M. Personalized time-aware tweets summarization. Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval. ACM, 2013; 513-522. DOI: 10.1145/2484028.2484052
29. Pang B., Lee L. Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval. 2008; 2(1). URL: http://www.cs.cornell.edu/home/llee/omsa/omsa.pdf
30. Cucerzan S., Brill E. Spelling correction as an iterative process that exploits the collective knowledge of web users. Proceedings of the Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2004; 293-300. URL: https://www.aclweb.org/anthology/W04-3238/

Прогнозирование тенденций развития и цен на рынке недвижимости с помощью предварительной обработки и анализа тональности текста

Повсеместный рост в области интеллектуального анализа текста, где интеллектуальный анализ социальных сетей играет значительную роль, является беспрецедентным. Это становится потенциальным источником изучения и анализа познаний людей, особенно с помощью предварительной обработки анализа тональности и текста. Анализ потенциального влияния настроений на реальные рынки недвижимости вызывает дискуссии ученых из области финансов, оценки и рыночной эффективности. Следовательно, это является весьма важной задачей для нашего исследования, которое не только обеспечивает подходящую платформу для подобных обсуждений, но и для всех, активно ищущих информацию о рынке недвижимости. Интеллектуальный анализ текста привлек внимание пользователей информации о рынке недвижимости, который находится на грани серьезных трансформаций. Доступность данных в таком гигантском объеме требует регулярного и критического анализа всей этой информации на предмет правдоподобия различных точек зрения. Огромные объемы текстов в социальных сетях онлайн можно использовать для эффективного извлечения релевантной информации о недвижимости. Поскольку интеллектуальный анализ текста играет важную и решающую роль в раскрытии этой идеи, необходимо тщательно изучить данные проблемы и их вклад в развитие анализа социальных сетей. В этой статье кратко рассмотрено текущее состояние предварительной обработки и анализа тональности текста для анализа рынка недвижимости. Особое внимание уделяется источникам и механизму изучения, доступным для исследователей и практиков в сфере недвижимости, также обсуждаются основные задачи интеллектуального анализа текста, представляющие интерес для широкого круга потребителей результатов исследования. Таким образом, основная цель этой статьи состоит в том, чтобы разъяснить и исследовать те области социальных сетей, которые широко доступны для предварительной обработки и анализа тональности текста о недвижимости для прогнозирования тенденций и цен на рынке недвижимости.

***Ключевые слова**: недвижимость, анализ тональности, социальные сети, интеллектуальный анализ текста, спам, CRF, n-грамм, энтропия*

ЛИТЕРАТУРА

1. Lifna C.S, Vijayalakshmi M. Identifying concept-drift in twitter streams // Procedia Computer Science. 2015. Vol. 45. Pp. 86–94. DOI: 10.1016/j.procs.2015.03.093
2. Goldhahn D., Eckart T., Quasthoff U. Building large monolingual dictionaries at the Leipzig Corpora collection : From 100 to 200 languages // Proceedings of the Eighth International Conference on Language Resources and Evaluation. 2012. Pp. 759–765. URL: http://www.lrec-conf.org/proceedings/lrec2012/pdf/327_Paper.pdf
3. Singh T., Kumari M. Role of text pre-processing in twitter sentiment analysis // Procedia Computer Science. 2016. Vol. 89. Pp. 549–554. DOI: 10.1016/j.procs.2016.06.095
4. Bhadane C., Dalal H., Doshi H. Sentiment analysis: Measuring opinions // Procedia Computer Science. 2015. Vol. 45. Pp. 808–814. DOI: 10.1016/j.procs.2015.03.159
5. Hamdan H., Bellot P., Bechet F. Lsislif: Feature extraction and label weighting for sentiment analysis in Twitter // Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015). Association for Computational Linguistics, 2015. Pp. 568–573. DOI: 10.18653/v1/S15-2095
6. Saleiro P., Rodrigues E.M., Soares C. FEUP at SemEval-2017 Task 5 : Predicting Sentiment Polarity and Intensity with Financial Word Embeddings. 2017. Pp. 904–908.
7. Vavliakis K.N., Symeonidis A.L., Mitkas P.A. Event identification in web social media through named entity recognition and topic modeling // Data & Knowledge Engineering. 2013. Vol. 88. Pp. 1–24. DOI: 10.1016/j.datak.2013.08.006
8. Zhou X., Chen L. Event detection over twitter social media streams // The VLDB Journal. 2014. Vol. 23. No. 3. Pp. 381–400. DOI: 10.1007/s00778-013-0320-3
9. Zhao Q., Mitra P. Event detection and visualization for social text streams // ICWSM 2007 — International Conference on Weblogs and Social Media. 2007. Pp. 26–28.
10. Aiello L.M., Petkos G., Martin C., Corney D., Papadopoulos S., Skraba R. et al. Sensing trending topics in twitter // IEEE Transactions on Multimedia. Institute of Electrical and Electronics Engineers, 2013. Vol. 15. No. 6. Pp. 1268–1282. DOI: 10.1109/TMM.2013.2265080
11. Lin J., Snow R., Morgan W. Smoothing techniques for adaptive online language models: Topic tracking in tweet streams // Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining — KDD '11. 2011. Pp. 422–429. DOI: 10.1145/2020408.2020476

12. Martinez L., Barranco M.J., Pérez L.P., Espinilla M. A knowledge based recommender system with multigranular linguistic information // International Journal of Computational Intelligence Systems. 2008.Vol.1.No.3.Pp.225–236.DOI:10.1080/18756891.2008.9727620
13. Quan C., Ren F. Unsupervised product feature extraction for feature-oriented opinion determination // Information Sciences. 2014. Vol. 272. Pp. 16–28. DOI: 10.1016/j.ins.2014.02.063
14. Croft W.B., Metzler D., Strohman T. Search engines: Information retrieval in practice. Pearson Education. 2015. 542 p.
15. Miner G., Delen D., Elder J., Fast A., Hill T., Nisbet R. The seven practice areas of text analytics. Elsevier. 2012. Pp. 29–41. URL: https://www.elderresearch.com/hubfs/Whitepaper_The_Seven_Practice_Areas_of_Text_Analytics_Chapter_2_Excerpt.pdf
16. Ghai D., Gera D., Jain N. A new approach to extract text from images based on DWT and K-means clustering // International Journal of Computational Intelligence Systems. 2016. Vol. 9. No. 5. Pp. 900–916. DOI: 10.1080/18756891.2016.1237189
17. Aggarwal C.C. Chapter 10. A survey of stream clustering algorithms // Data Clusterin. 2013. Pp. 229–252. DOI: 10.1201/9781315373515-10
18. Li X., Yu W. Fast support vector machine classification for large data sets // International Journal of Computational Intelligence Systems 2013. Vol. 7. No. 2. Pp. 197–212. DOI: 10.1080/18756891.2013.868148
19. Srivastava J., Desikan P., Kumar V. Web mining — Concepts, applications and research. 2002. Pp. 51–71. URL: https://www-users.cs.umn.edu/~desi0016/publications/wmo.pdf
20. Popowich F. Using text mining and natural language processing for health care claims processing // ACM SIGKDD Explorations Newsletter. 2005. Vol. 7. No. 1. Pp. 59–66. DOI: 10.1145/1089815.1089824
21. Gelfand B., Wulfekuhler M., Punch W. Automated concept extraction from plain text. AAAI Technical Report WS-98-05. 1998. URL: https://www.aaai.org/Papers/Workshops/1998/WS-98-05/WS98-05-003.pdf
22. Kumar J.P., Govindarajulu P. Near-duplicate web page detection: An efficient approach using clustering, sentence feature and fingerprinting // International Journal of Computational Intelligence Systems. 2013. Vol. 6. No. 1. Pp. 1–13. DOI: 10.1080/18756891.2013.752657
23. Bhuta S., Doshi U. A review of techniques for sentiment analysis of twitter data // 2014 International Conference on Issues and Challenges in Intelligent Computing Techniques (ICICT). IEEE, 2014. Pp. 583–591. DOI: 10.1109/ICICICT.2014.6781346
24. Peetz M.-H. Time-aware online reputation analysis. 2015. URL: https://pure.uva.nl/ws/files/2550354/157890_peetz_thesis_complete.pdf
25. Ren Z. Monitoring social media: Summarization, classification and recommendation : PhD thesis. 2016. URL: https://hdl.handle.net/11245/1.541961
26. Han B. Improving the utility of social media with natural language processing : PhD thesis. 2014. URL: http://hdl.handle.net/11343/41029
27. Aichner T., Jacob F. Measuring the degree of corporate social media use // International Journal of Market Research. 2015. Vol. 57. No. 2. Pp. 257–276. DOI: 10.2501/IJMR-2015-018
28. Ren Z., Liang S., Meij E., de Rijke M. Personalized time-aware tweets summarization // Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval. ACM, 2013. Pp. 513–522. DOI: 10.1145/2484028.2484052
29. Pang B., Lee L. Opinion mining and sentiment analysis // Foundations and Trends in Information Retrieval. 2008. Vol. 2. No. 1. URL: http://www.cs.cornell.edu/home/llee/omsa/omsa.pdf
30. Cucerzan S., Brill E. Spelling correction as an iterative process that exploits the collective knowledge of web users // Proceedings of the Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2004. Pp. 293–300. URL: https://www.aclweb.org/anthology/W04-3238/

Об авторах: **Синяк Николай Георгиевич** — кандидат экономических наук, профессор кафедры менеджмента и экономики; **Частный институт управления и предпринимательства**; Беларусь, 220086, г. Минск, ул. Славинского, д. 1, корп. 3; SPIN-код: 4400-2224; Scopus: 55952470200; ResearcherID: K-4838-2015; ORCID: 0000-0002-1688-9268; Google Scholar: 9wEDUrMAAAAJ; siniakn@gmail.com;

Тажиндер Сингх — доцент кафедры компьютерных наук и инженерии, **Институт инженерии и технологий Сант Лонговал**; Индия, 148106, г. Пенджаб, Сангрур, Административный блок SLIET, SLIET Rd; Google Scholar: nn4t8UMAAAAJ&hl; nith2k14@gmail.com;

Мадху Кумари Джаглан — кандидат наук, доцент, доцент кафедры компьютерных наук и информационных технологий; **Национальный технологический институт Хамирпура**; Индия, 177005, Химачал-Прадеш, г. Хамирпура; ORCID: 0000-0003-3203-2579; Madhu.jaglan@gmail.com;

Козловский Виталий Владимирович — доктор экономических наук, главный научный сотрудник; **Научно-исследовательский экономический институт Министерства экономики Республики Беларусь (НИЭИ Минэкономики)**; Беларусь, 220086, г. Минск, ул. Славинского, д. 1, корп. 1; SPIN-код: 4354-7264, ResearcherID: ABI-8122-2020, ORCID: 000-0002-9194-6170; Vital_kozlovsky@mail.ru.

For citation: Sinyak N.G., Tajinder S., Madhu K.J., Kozlovskiy V.V. Predicting real estate market trends and value using pre-processing and sentiment text mining analysis. *Real estate: economics, management*. 2021; 1:35-43.

Для цитирования: *Синяк Н.Г., Тажиндер С., Мадху К.Д., Козловский В.В.* Predicting real estate market trends and value using pre-processing and sentiment text mining analysis // Недвижимость: экономика, управление. 2021. № 1. С. 35–43.

Павлова Л.И. Москва. Гостиница «Метрополь». Рисунок, тушь, перо