information

MDPI

Article

Lift Charts-Based Binary Classification in Unsupervised Setting for Concept-Based Retrieval of Emotionally Annotated Images from Affective Multimedia Databases

Marko Horvat [1,*], Alan Jović [2] and Danko Ivošević [3]

[1] Department of Applied Computing, Faculty of Electrical Engineering and Computing, University of Zagreb, Unska 3, HR-10000 Zagreb, Croatia

[2] Department of Electronics, Microelectronics, Computer and Intelligent Systems, Faculty of Electrical Engineering and Computing, University of Zagreb, Unska 3, HR-10000 Zagreb, Croatia; alan.jovic@fer.hr

[3] Department of Computer Science and Information Technology, Zagreb University of Applied Sciences, Vrbik 8, HR-10000 Zagreb, Croatia; danko.ivosevic@tvz.hr

* Correspondence: marko.horvat3@fer.hr; Tel.:+385-1-6129-531

Received: 20 June 2020; Accepted: 1 September 2020; Published: 3 September 2020

Abstract: Evaluation of document classification is straightforward if complete information on the documents' true categories exists. In this case, the rank of each document can be accurately determined and evaluated. However, in an unsupervised setting, where the exact document category is not available, lift charts become an advantageous method for evaluation of the retrieval quality and categorization of ranked documents. We introduce lift charts as binary classifiers of ranked documents and explain how to apply them to the concept-based retrieval of emotionally annotated images as one of the possible retrieval methods for this application. Furthermore, we describe affective multimedia databases on a representative example of the International Affective Picture System (IAPS) dataset, their applications, advantages, and deficiencies, and explain how lift charts may be used as a helpful method for document retrieval in this domain. Optimization of lift charts for recall and precision is also described. A typical scenario of document retrieval is presented on a set of 800 affective pictures labeled with an unsupervised glossary. In the lift charts-based retrieval using the approximate matching method, the highest attained accuracy, precision, and recall were 51.06%, 47.41%, 95.89%, and 81.83%, 99.70%, 33.56%, when optimized for recall and precision, respectively.

Keywords: image classification; image retrieval; concept based retrieval; affective computing; performance evaluation; lift charts

1. Introduction

The ever-growing size and complexity of unstructured information available on the World Wide Web continuously motivate computer science researchers in the development of more useful data description models and retrieval methods. Classification is a constituent part of every image retrieval system. Efficient classifiers, which benefit from high information retrieval performance and correctness metrics, such as accuracy, precision, fall-out, F-measure, mean average precision, or discounted cumulative gain [1,2] will remove irrelevant results in the returned dataset and provide users only with the information they actually requested. Because of the high data throughput requirements in online document retrieval systems, classification speed is also very important for positive user experience. Therefore, it is imperative to identify the classification models with the optimal set of characteristics.

Lift charts are typically used for the evaluation of machine learning models and the comparison of one model's performance to another [3,4]. As a type of diagram, a lift chart graphically represents and

quantifies the improvement that a classifier provides when compared against a different classifier or a random guess [3,4]. Lift charts have two main applications. Firstly, they are a well-known method for high-quality evaluation of classification models used in data mining and machine learning [5]. With a lift chart, it is possible to compare the accuracy of predictions for multiple models that have the same predictable attribute. Furthermore, lift charts may be used to assess the accuracy of prediction either for a single outcome (a single value of the predictable attribute) or for all outcomes (all values of the specified attribute). Assuming the previous ranking of classification results (e.g., established with an appropriate relatedness measure between a query and document descriptors), the construction of a lift chart is algorithmically simple, with linear time complexity.

The second application of lift charts, which is very important and often overlooked, is in the classification of ranked retrieval results. Such results may be optimized for either precision or recall. This type of lift chart application is the subject of the paper. An example is provided in Figure 1 to improve clarity. In this flexible approach, it is possible to arbitrarily increase the ratio of correctly classified items without prior knowledge of the item classes at the expense of the end sample size. Although this optimization method may be equally applied to concept-based [6] or content-based [7] image retrieval for determining category boundaries, we will discuss in detail only the former implementation for the sake of brevity and conciseness.

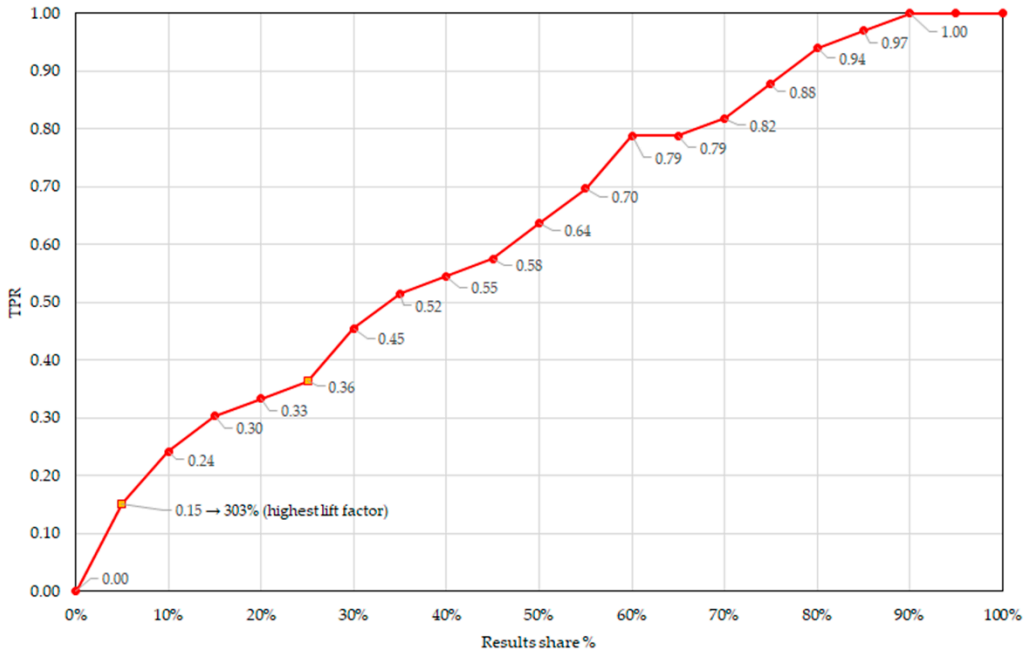

Figure 1. An example of a realistic lift chart from the evaluation experiment described in Section 5. The dataset was queried with two keywords and instances were ranked by approximate lexical similarity. Optimal precision was achieved for the first 5% of emotionally annotated pictures in the set, while the optimal recall (>90% of true positive examples) was ensured by using the top 80% ranked results.

The application of lift charts as classifiers is specifically directed towards affective multimedia databases that are used in experimentation with attention and emotion. These multimedia affective databases are a frequently used tool with numerous applications in the fields of psychology and neuroscience. Emotionally annotated multimedia are images, sounds, video, music and text documents with annotated semantic and emotional content. They are stored in affective multimedia databases. Apart from digital objects, these databases contain meta-data about their high-level semantics and the expected emotion that will be induced in a subject when exposed to a contained document. Two important features distinguish affective multimedia databases from other multimedia repositories: (1) the purpose of the multimedia documents and (2) the emotional representation of the multimedia

documents [8]. In our approach, binary classifiers based on lift charts are applied to improve precision and recall in text-based concept retrieval from these databases. Lit charts are a novel but by no means universally optimal retrieval tool for affective documents. Other classifiers suitable for multimedia retrieval may also be used for the retrieval of emotionally annotated images if they take into account different emotional and semantic models in the databases.

The remainder of the paper is organized as follows: in the next section, an outline of affective multimedia databases is given—what they are, how they are structured and how they are commonly used. After that, in Section 3, the problem of concept-based image retrieval and binary classification of images, where lift charts can be successfully applied, is formally defined. Lift charts and their properties are described in Section 4. The evaluation of lift charts on a concrete image retrieval task from the International Affective Picture System (IAPS) affective picture database is presented in Section 5. Advantages and shortcomings of lift charts as classifiers and methods for evaluation of classifiers, and how the performance of lift charts in the described setting may be improved, are discussed in Section 6. Finally, conclusions are presented in the final section, at the end of the paper.

2. Affective Multimedia Databases: Definition, Architecture and Usage

Affective picture databases are most frequently employed in the controlled stimulation of emotional reactions for experimentation in cognitive sciences, psychology, neuroscience, and different interdisciplinary studies, such as human–computer interaction (HCI) [9–11]. Simply put, they are experts' tools designed for intentionally provoking targeted emotional reactions in exposed subjects. They have many practical uses related to research in perception, memory, attention, reasoning, and, of course, emotion. Many such databases have so far been developed. The most popular ones, which are free for use by researchers, are: the International Affective Picture System (IAPS) [9,10], the Nencki Affective Picture System (NAPS) [12] (with its extensions NAPS Basic Emotions [13] and NAPS Erotic Subset [14]), the Geneva Affective Picture Database (GAPED) [15], the Open Library of Affective Foods (OLAF) [16], the DIsgust-RelaTed-Images (DIRTI) [17], the Set of Fear-Inducing Pictures (SFIP) [18], the Open Affective Standardized Image Set (OASIS) [19], and the most recent one, the Children-Rated Subset to the NAPS [20]. In addition, a recent list of affective picture databases with different emotionally-annotated multimedia formats and exemplars of the research conducted with these database is available in [8]. An example of affective pictures from the OASIS database, with semantic tags included, is shown in Figure 2. The OASIS employs similar emotional and semantic models to the IAPS database.

Figure 2. An example of 10 affective pictures with their tags from the Open Affective Standardized Image Set (OASIS) database [19]. Pictures with high arousal and valence for inducing the emotion of happiness (top row), and high arousal and negative valence for stimulation of fear (bottom row). Reproduced with permission from Kurdi, B.; Lozano, S.; Banaji, M.R., Introducing the open affective standardized image set (OASIS); published by Springer, 2017.

New affective databases for provoking emotional responses are being continuously developed. This trend is encouraged by the growth in the variety and complexity of, and in demands for intensity and precision in, the experimentation by the researchers. Although they are invaluable tools in their field of practice, the affective multimedia databases have two important drawbacks that can be at least alleviated, if not completely eliminated, by utilizing methods from computer science and information retrieval. These drawbacks are demanding and time-consuming construction and retrieval of affective multimedia documents. Both are related to inefficient search of the databases, which is caused by their rudimentary and inadequate semantic representation model.

It is hard to define a specific unitary structure of affective multimedia databases, since there is no accepted standard for their construction. However, although they are different, some important common, distinctive features may be established [8]. Most importantly, and in the context of this paper, document semantic annotations use a sparse bag-of-words model. In the affective databases, a single multimedia stimulus is tagged with an unsupervised glossary. Frequently, a document is tagged with a single free-text keyword, and different tags are used to describe the same concept. For example, a picture showing an attack dog in the IAPS database could be tagged as "dog", "attack", "AttackDog", or "attack_dog", etc. Synonym tags such as "canine" or "hound" are also used. Furthermore, semantic relations between different concepts are undefined. The lexicon itself does not implement semantic similarity measures and there are no criteria to estimate relatedness between concepts. For example, in such a model, it is difficult to determine that "dog" and "giraffe" are closer to each other than "dog" and "door". This is a huge flaw in the document retrieval process because a search query must match the keywords only lexically. In this setting, a more semantically meaningful interpretation of the query and annotating tags is not possible. The inadequate semantic descriptors result in three negative effects that impair information retrieval: (1) low recall, (2) low precision and high recall, or (3) vocabulary mismatch. Moreover, affective multimedia databases do not contain their own hierarchical model of semantic labels and do not reuse external knowledge bases.

Unlike semantics, the description of multimedia affect is much more efficient and standardized across the currently published affective multimedia databases. The two most common models of emotion are the pleasure-arousal-dominance (PAD) dimensional [21] and discrete models [22].

2.1. The Please-Arousal-Dominance Emotion Model

The PAD dimensional model is defined with three emotion dimensions: valence (Val), arousal (Ar) and dominance (Dom). Often, only the first two dimensions are used, because dominance is the least informative measure of the elicited affect [23]. All dimensions are described with continuous variables in bounded but closed intervals from 1.0 to 9.0 as $al \in [1,9] \in Val, ar \in [1,9] \in Ar$, and $dom \in [1,9] \in Dom$. These three affective dimensions have orthogonal meanings in psychology [21,23]. Valence describes positivity and negativity of stimuli, while arousal specifies the intensity or energy level of a stimulus, and dominance is the controlling and dominant nature of the emotion. Most often, dominance is omitted, since it was experimentally established that it does not significantly contribute to discrimination of stimuli. Thus, affective values in multimedia are commonly described as coordinates in a 2D Cartesian space with the valence dimension representing the x-axis and arousal the y-axis. In such a model, the distance between two emotions can be simply expressed as a Euclidian distance between two points in the coordinate system. If $pic_1(val_1, ar_1)$ and $pic_2(val_2, ar_2)$ are two pictures with dimensional emotion coordinates val_1, val_2 and ar_1, ar_2 representing their respective valence and arousal values, then the emotion distance $d_{emo}^{dimensional}$ between pic_1 and pic_2 in the dimensional model is Equation (1):

$$d_{emo}^{dimensional} = \left|pic_1, pic_2\right| = \sqrt{(val_1 - val_2)^2 + (ar_1 - ar_2)^2} \quad (1)$$

2.2. The Discrete Emotion Model

The discrete emotion model, also called emotion norms, is defined around an immutable set of distinct emotional categories, where the intensity of each of these basic emotions is a unified pair.

The intensity is represented with a real value between 0.0 and 1.0, where 1.0 represents the maximum affective intensity. For example, the widely used Ekman's model, also called "The Big Six", defines six different basic emotions: happiness, disgust, anger, fear, sadness, and surprise. However, researchers do not unanimously agree on the right number of categories, and affective multimedia databases were built on models with different numbers of discrete emotions. This makes the transformation of a stimulus from one database to another, or unification of datasets from dissimilar databases, difficult. In the discrete model difference between two affective pictures is a vector in n-dimensional space. If $pic_1(be_{11}, be_{12}, \ldots, be_{1n})$ and $pic_2(be_{21}, be_{22}, \ldots, be_{2n})$ are two pictures with intensities of n discrete emotions $be_{11}, be_{12}, \ldots, be_{1n}$ and $be_{21}, be_{22}, \ldots, be_{2n}$, then the emotion distance $d_{emo}^{discrete}$ between pic_1 and pic_2 in the discrete model is given with Equation (2):

$$d_{emo}^{discrete} = \|pic_1, pic_2\| = \sqrt{\sum_{i=1}^{n} (be_{1i} - be_{2i})^2} \quad (2)$$

where i represents the ordinal number or index of some emotion category.

The most comprehensive effort into the systematization of emotional models for usage in computer data processing systems is represented by the EmotionML standard recommendation developed under the umbrella of the W3C Consortium [24]. The Emotion Markup Language is written in XML and serves as an annotation language for multimedia. It currently offers the most sophisticated emotion glossary, which includes five category vocabularies, four dimension vocabularies, three appraisal vocabularies, and one action tendency vocabulary. In studies from psychology and neuroscience, additional models, apart from the PAD dimensional and the discrete ones, have been proposed, but these theories have not been applied to the annotation of affective multimedia (see [24] for more information).

The semantics of affective multimedia databases was manually annotated by domain experts, while the emotional content was estimated and verified experimentally. The elicited emotion values are acquired with rigidly controlled experiments in which participants express, most commonly in the form of a questionnaire, their subjective beliefs in elicited emotions. For instance, in the construction of the latest version of the IAPS database, over 100,000 standardized self-reports were collected and statistically analyzed to establish statistical distributions of emotional dimensions for 1195 pictures [9,10]. Hence, emotional reactions of around one hundred participants on average have been aggregated per picture. A similar approach was used in other databases, while only some of them additionally have recordings of subjects' physiological signals in baseline and excited states.

3. Overview of Concept-Based Image Retrieval with Boolean Classification

In concept-based, also known as description-based or text-based, image retrieval, binary or binomial classification is an essential step in the selection of true results [25]. In this process, lift charts can be used in two different ways: (1) to determine the cutoff rank of retrieved items, and (2) to assess the quality of the retrieval itself.

In concept-based document retrieval systems, as well as in image retrieval systems as their subtype, a new search is started by entering a query that constraints the items that should be retrieved. The query must be translated into the set of concepts that represents a set of documents being searched for adequately well [1,2]. The formality and expressivity of the concept set depend on the chosen knowledge representation model. Transformation of text to concepts may be trivial in the case where it is performed as a series of search tags or keywords. The search may greatly benefit if an underlying knowledge base is used that defines concept labels, their properties, and relationships between them [7].

In the next step, the query concepts must be brought into a functional relationship and quantitatively compared with descriptions of images stored in the multimedia repository. With a technique called "query by example", the concepts are not entered directly, but fetched from a description of an archetypical image, which was uploaded into the system instead of the query [7]. In text-based systems, describing concepts must be provided together with the example

image itself. This is usually accomplished by letting users choose one or more similar images already in the repository [1,2].

After a search query has been entered, in the next step of the retrieval process, each concept in the search text is compared to every concept c_i in descriptions of images stored in the repository. A suitable measure is then used to rank images according to their relatedness or similarity to the search query. The similarity measure is a function $sim\left(c_i, c_j\right) \in [0,1] : \ c_i, c_j \subset C$, which takes two concepts from the set of all concepts C as its domain and provides a real number in a closed interval between 0.0 and 1.0 as a measure of closeness between the two concepts c_i and c_j. Similarity measures between pairs of concepts in the query $q_i, q_j \subset C$ and image description $d_i, d_j \subset C$ are combined with an aggregation function. The aggregation may be any suitable function, but usually it is a sum or product of individual similarity assessments [26]. If items in the repository are described with labels from a knowledge base, then it is possible to use concept distance measures, which are more formal and semantically more meaningful than lexical measures. Examples of both sets of measures can be found in [1,2]. Apart from the expressiveness of image description, the choice of the similarity and aggregation functions is crucial for the quality of the retrieval.

The results, i.e., retrieved documents, are represented as an ordered list $a_1, a_2, \ldots, a_n$. Each image is assigned with a rank $r \in \mathbb{Z}$, which is the sequence number of the image in the returned list. Consequently, if M items are being returned, the first item a_1 has the rank $r_1 = 1$, the second one $r_2 = 2$, and the rank of the last item a_n is $r_n = |M|$. For a search to be considered successful, documents closer to the posed query should appear first, i.e., near the beginning of the returned results, and conversely, less related documents should appear more towards the end of the list. In other words, items more relevant to the query have a lower rank and those less relevant have a higher rank number. Ideally, the similarity of items in the returned list should be a monotonically decreasing and continuous function. As users browse through the list, from the first document towards the last, they expect their relevancy to continuously decrease. It would be counterintuitive and indeed harmful if items being sought are located at the end of the list. Such items may be unintentionally overlooked. Therefore, in an ideal series Equation (3) is valid:

$$\forall s_i \geq s_{i+1} \ : i = 1, 2, \cdots, |M| - 1 \tag{3}$$

where $s_i \in \mathbb{R}$ is the similarity of item a_i. If a large number of documents are stored in the repository, it would be costly to retrieve all of them at once. In such circumstances, only the best documents should be presented, while others are disregarded. The decision boundary is set at a rank cutoff $r_c \in [0, |M|]$, where items $a_i : i \leq r_c$ will be displayed to users and $a_j : j > r_c$ will not. Preferably, the cutoff value must be chosen so that precision and recall are maximized in order to display as many relevant documents in the repository as possible. Finally, after the documents have been classified and ranked, they may be extracted from the repository with their textual descriptions and presented to the user in the specific order consistent with their appointed rank.

3.1. Evaluation Metrics

When dealing with two-class (i.e., binary) classification problems in picture retrieval, only one class contains objects that match the search parameters and will be presented to users, while objects in the other class are discarded as irrelevant. The first class is usually labeled "positive" or "true" and the other "negative" or "false".

The dataset always consists of P positive and N negative examples. The job of a classifier is to assign a class to each of them. In realistic settings, some of the assignments will inevitably be wrong. To assess the classification outcome, we have to count the number of: (1) true positives (TP), (2) true negatives (TN), (3) false positives (FP) as negative, but wrongly classified as positive, and (4) false negatives (FN) as actually positive examples that are classified as negative. By definition, positive and negative results are related as in Equations (4) and (5):

$$P = TP + FN \tag{4}$$

and

$$N = TN + FP. \tag{5}$$

In this regard, any classifier, either realistic or optimal, always assigns $TP + FP$ examples to the positive class and $TN + FN$ examples to the negative class. However, in practice, information retrieval systems do make mistakes: false negatives are retrieved and displayed, although they should be rejected, while false positive images are not presented to a user by a retrieval system, although they should be [27,28]. These classification errors are called Type I and Type II errors, respectively.

The most widely used measures of classifier performance and correctness are accuracy, precision, recall, fall-out and F-measure. Precision and accuracy are often used to measure the quality of binary classifiers. Precision, or positive predictive value (PPV) as it is also called, is defined in Equation (6) as the proportion of accurately classified examples in a set of positively classified examples [27,28]:

$$PPV = \frac{TP}{TP + FP} \tag{6}$$

Recall (R), also named sensitivity, hit rate, and true positive rate (TPR), is the proportion of accurately classified examples in the set of all positive examples [27,28] as in Equation (7):

$$R = TPR = \frac{TP}{TP + FN} \tag{7}$$

Precision may also be defined as the probability that a retrieved document is relevant, while recall is the probability that a relevant document is retrieved in a search. Additional performance measures for lift charts can be defined, as described in the next section. If classification is binary or binomial, i.e., with only two classes, the term sensitivity is often used instead of recall. In this case, FP rate ($FPrate$) is designated as FPR, and TP rate ($TPrate$) as TPR [29]. Fall-out is the proportion of non-relevant documents that are retrieved out of all non-relevant documents, while F-measure (also F_1-score or F-score) includes precision and recall as its weighted harmonic mean. A lower result for fall-out is better.

4. Binary Classification with Lift Charts

Lift charts are a type of charts, such as the receiver operating characteristic (ROC) curves and precision-recall curves, which are often utilized in machine learning for visualization and evaluation of classification models. They can be especially useful in cases where the number of false positive observations is unavailable and, subsequently, more common machine learning evaluation methods, such as the ROC curves, cannot be constructed. Lift charts are primarily used as a tool for observing the improvement that a classifier makes against a random guess. In previously published literature they are described in this respect [3,30,31].

However, the second use of lift charts—which is in the focus of this paper—is in the binary classification of ranked results. In this application, lift charts can be optimized to increase retrieval precision or recall. The difference in optimization is achieved with the choice of rank cutoff. In optimization for precision, the cutoff is commonly set to a lower rank, while if classification is adjusted for recall (i.e., better sensitivity), the cutoff is set at a higher rank. The exact choice of rank depends on the data and the desired classification performance.

In the document retrieval experiment, the lift charts have been demonstrated to be a helpful method for binomial classification in situations where ground truth annotations are inadequate or unattainable and the precise rank of affective multimedia documents cannot be determined. Ground truth annotations or labels are very important in concept-based retrieval, as they represent the true description of documents. In fact, they specify an objective and complete knowledge about document content in description-based retrieval paradigms. The annotations can be represented by any adequately expressive model, such as keywords from supervised or controlled vocabularies as in the

bag-of-words model [1,2]. Ground truth labels of images are always provided by a trusted authority. The labels are added either by a human domain expert or automatically by image analysis, depending on the complexity of the problem [32]. However, in some information retrieval applications, the labels are not available or cannot be correctly defined. In such circumstances, both the true category and rank of retrieved results cannot be known accurately and, subsequently, the search performance deteriorates. In these cases, the capability to assign documents to the correct class becomes more important than the effectiveness of finding the correct document rank [1,2]. The rationale is that if the rank is incorrect and document order in the results list is inaccurate, at least they will be correctly classified and presented to a user. In other words, it is better to retrieve a correct document albeit with an incorrect position in the results list, than to not retrieve it at all. The lift charts can assist in improving precision and recall in such settings.

It should be noted that the term "profit charts" is sometimes used in the literature instead of "lift charts" (for example in [33]). However, the two are not completely identical. A profit chart contains the same information as a lift chart, but also presents the estimated increase in profit that is related to a specific model [3].

Formally, in document retrieval, a lift chart is a two-dimensional (2D) graph with its x-axis representing the ranked number of results and its y-axis showing the true positive (TP) measure. Such a lift chart may be defined as in Equation (8):

$$x = t,\ y = TP(t) \tag{8}$$

where N is the total number of documents being classified, $t \in [0, N]$ is the ranked document's ordinal number, and $TP(t) \in [0, N]$ is the true positive value at position t.

However, for better convenience, relative ratios are often used for the definition of chart axes instead of total numbers. In this approach, the x-axis represents the proportion of ranked results (%) and the y-axis represents the true positive rate (TPR), which is the ratio of correctly classified instances and total number of documents in the retrieved set, as was explained in Section 3.1. Consequently, the lift chart function can be defined as in Equation (9):

$$x = \frac{t}{N},\ y = TPR(t) \tag{9}$$

where $x \in [0.0,\ 1.0]$ represents the proportion of the result and $TPR(t) \in [0.0,\ 1.0]$ is the value of the true positive rate at ranked position t.

Therefore, a lift chart is created by calculating $TP(t)$ or $TPR(t)$ for specific values of t. In practice, lift charts are not smooth but stepwise, i.e., each point on the graph defines a column (a step or an increment) with the point in its upper left corner. To facilitate and precipitate the plotting process, usually only representative values of t are considered in discrete increments, such as each 5% (t/N = 0.0, 0.05, 0.1, ... , 0.95, 1.0) or 10% (t/N = 0.0, 0.1, 0.2, ... , 0.9, 1.0) data segments. The convex curve obtained by drawing the given points is called a lift chart. The coarseness of charts' shapes is directly proportional to their increment size. Smaller steps, as the difference between two consecutive values of t, will have the effect of a smoother curve shape, while larger steps will produce wider discrete columns and a coarser or step-like hull. Note that the curve can be at most calculated for each classified item, which is not recommended if the retrieved set is very large.

An ideal classifier model will have $TPR(t_0) = 1.0$ for $t_0 = 0$, because a perfect retrieval system will provide only TP documents to users, correctly disregarding TN, FN, and FP instances. The corresponding lift chart is a vertical line along the left edge and a horizontal line on the top edge of the plot area. On the other hand, the lift chart of a naïve classifier that categorizes documents just by random guessing is a straight diagonal line from the lower left to the upper right corner. Such a lift chart is defined by two points, $TPR(t_0) = 0.0$, $t_0 = 0$ and $TPR(t_1) = 1.0$, $t_1 = 1.0$.

In binary settings, a random classifier classifies 50% of pictures to the "*True*" category and 50% to the "*False*" category. As such, it has the worst performance of all possible classifiers and is used as a preferable reference for the evaluation of machine learning models. Thus, it is commonly assumed that the performance of any actual classification model is better than the classification performed simply by chance and worse than that of an ideal classifier.

Three lift charts of a realistic classifier from the experiment in this paper, for ideal and random classifiers with an increase of 5% in t, are shown in Figure 3.

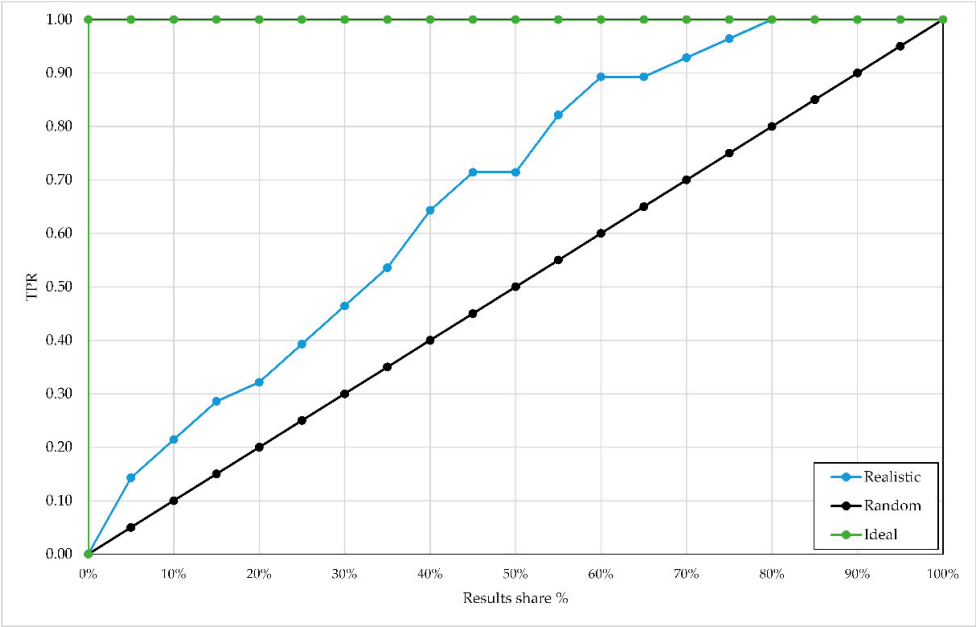

Figure 3. Three representative lift charts plotted in 5% increments: (1) realistic stepwise (blue line), (2) random guess—referent (black), (3) ideal (green). True positive rate is placed on the y-axis and share of total results on the x-axis.

The improvement of a model is evaluated in terms of a numerical measure called the lift score or, more commonly, the lift. By comparing the lift for all portions of a dataset and for different classifiers, it is possible to determine which model is better, or even optimal, and which percentage of the cases in the dataset would benefit from applying the model's predictions. Classifiers may have similar or even identical performance for some values of t/N, making the lift negligible or very small, but, at the same time, their performance can be considerably different in other sections. Therefore, the lift must be determined for all steps to establish the maximum score.

Formally, the lift is a measure of predictive model effectiveness calculated as the ratio between the results obtained with and without the predictive model. The lift is commonly expressed in relative terms as a percentage where, for example, a lift of 100% implies a double improvement in predictiveness compared to a referent model, a lift of 200% a triple improvement, and so on.

The area under curve (AUC), or more precisely, the area under lift chart A_{lift} can be used as a measure of classification quality. Since, in practice, lift charts are not smooth, but stepwise, the sum of all discrete columns in a stepwise lift chart is equal to A_{lift}.

Using integration, it can be easily shown, as in Equation (10), that the random classifier has an area under the curve equal to:

$$A_{lift} = \frac{1}{P+N}\left(\frac{P^2}{2} + PN{\cdot}A_{ROC}\right);\ \ A_{lift} \in [0, P] \tag{10}$$

where A_{ROC} is area under the curve of the same classifier [5]. It is defined in Equation (11):

$$A_{ROC} = \frac{1}{PN} \int_0^N TP \, dFP; \;\; A_{ROC} \in [0, 1] \tag{11}$$

The random classifier has $A_{lift} = P/2$ and a perfect classifier has $A_{lift} = P$. Therefore, the A_{lift} of actual classifiers lies between $P/2$ and P. Since A_{lift} always depends on the P to N ratio, if $P \ll N$, then it is possible to use the approximation $A_{lift} \cong A_{ROC} \cdot P$. Moreover, it should be noted that a random classifier has $A_{ROC} = 0.5$, while a perfect classifier has $A_{ROC} = 1$. Classifiers used in practice should therefore be somewhere in between, and preferably have an A_{ROC} close to the value of 1.

AUCs of the lift charts from Figure 3 are portrayed in Figure 4. In this case, with the relative proportion in the dataset on the x-axis, numerical integration for the realistic classifier gives $A_{lift} = 0.6714$. This is 1.3428 times more than for the random guess ($A_{lift} = 0.5$) and 1.4894 times less than for a perfect classifier ($A_{lift} = 1.0$).

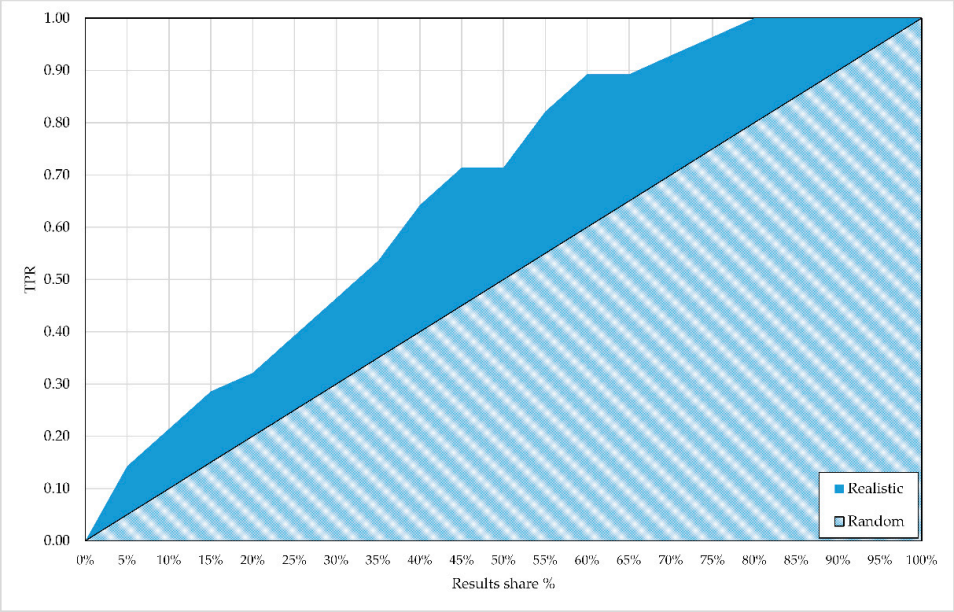

Figure 4. Areas under lift chart A_{lift} for a realistic (blue area) and random classifier (shaded blue area). The A_{lift} of the realistic classifier includes that of the random classifier. As in the previous figure, true positive rate is represented on the y-axis and share of total results is on the x-axis.

5. Evaluation of Affective Image Retrieval with Lift Charts as Binomial Classifiers in Ranking

In the experiment, binomial classification of images was evaluated with lift charts to provide a real-life example and indicate benefits and drawbacks of their utilization in image retrieval. The experiment demonstrated the application of lift charts as explained in Section 4. For this, a dataset $\mathbf{D}$ consisting of $N = 800$ pictures was taken from the IAPS corpora [9,10]. This database was chosen because it is the most frequently utilized one, and also one of the largest affective multimedia databases [8]. Moreover, its architecture may be considered typical for this type of datasets. Other repositories have comparable semantic models with unsupervised annotation glossaries and sparse tagging. The selected pictures are visually unambiguous with easily comprehensible content. The content of each picture was described with one keyword $w_i \in \mathbf{W}$ from an unsupervised glossary $\mathbf{W}$. The glossary contained 387 different keywords. The set of all queries $\mathbf{Q}$ applied in the evaluation consisted of different keywords taken from the annotation glossary ($\mathbf{Q} \subseteq \mathbf{W}$). For every query $q_i \in \mathbf{Q}$, a subset of documents $\dot{\mathbf{D}} \subset \mathbf{D}$ with $\left|\dot{\mathbf{D}}\right| = 100$ pictures was randomly preselected and then classified using lexical relatedness measures [34]. Each subset $\dot{\mathbf{D}}$ was queried three times: with one, two, or three different words.

Only the terms from **D** were used as queries, which guaranteed that the retrieved sets were nonempty. The setup for the experiment was reused from a previous research [34] but with the expanded dataset **D**, which also resulted in a different glossary **W**, queries **Q** and assortment of all preselected subsets $\dot{\mathbf{D}}$. Compared to [35], the additional images in the dataset provide greater credibility in the results of this study.

Two lexical relatedness measures were used for ranking: naïve string matching and the Levenshtein distance (i.e., edit distance) [34]. Each subset $\dot{\mathbf{D}}$ was classified once for each relatedness algorithm, with rank cutoff independently optimized twice: first for precision and then for recall. The goal of classification for precision was to maximize the fraction of pictures relevant to posited queries. Classification for recall tried the opposite: to maximize the share of relevant pictures in $\dot{\mathbf{D}}$. In practice, the two optimization methods are contradictory: the former results in high accuracy and a small number of retrieved samples, and the latter optimization returns a large share of samples, but with considerably lower accuracy.

5.1. Similarity Measures and Ranking

The two ranking algorithms used in the experiment assigned a similarity score (i.e., measure of lexical relatedness) between two text labels, a and b. This measure $\mathrm{rel}(a, b) \in [0,\ 1]$ with $a, b \in \mathbf{W}$ has the following properties in Equation (12):

$$\begin{array}{l} \mathrm{rel}(a, b) = 1,\ \mathrm{x} = \mathrm{y} \\ \mathrm{rel}(a, b) < 1,\ x \neq y. \end{array} \tag{12}$$

The string matching algorithm represents the most unassuming rule that merely checks if a specific series of characters is a part of another character series. The output of this Boolean model is binary (e.g., gives 0 or 1 as its output); either two terms do not match at all or they match completely. The string matching and Levenshtein algorithms were chosen for the experiment since they are typical and easy to understand representatives of two different types of lexical searching algorithms—exact and approximate, respectively. The first algorithm does not allow errors in search terms, unlike the latter, which permits users to search with wrong or imprecise terms and still get at least partially correct results. The naïve matching algorithm is elementary and can be easily implemented. On the other hand, Levenshtein is a popular approximate search algorithm that represents an entire group of edit distance lexical metrics [34]. Aside from string matching and Levenshtein distance, many concept-based retrieval methods exist in the literature that improve on and even surpass these two basic algorithms [36]. Therefore, the results from string matching and Levensthein distance algorithms can be regarded as good indicators of the minimum performance that lift charts can provide in binary classification for concept-based image retrieval.

As an example, the experiment picture 1019.jpg is tagged with k = "Snake" and the 1-word search query was q_1 = "Serpent". The naïve method (i.e., exact matching) resulted in $\mathrm{rel}_{exact}(k, q_1) = 0$, since k and q_1 were not completely equal. For the same example, the Levenshtein measure, as an approximate lexical matching algorithm, gave $\mathrm{rel}_{approx}(k, q_1) = 0.2$. For a 2-word query $q_2 = \left(q_2^1 \vee q_2^2\right)$, where q_2^1 = "Snake" and q_2^2 = "Serpent", and for the same picture keyword k, aggregation of separate relevance scores was necessary. In the experiment, arithmetic mean, as the aggregation function, was selected among a range of other relevance score combination methods [27]. Thus, for the 2-word query q_2 the exact matching is defined in Equation (13):

$$\mathrm{rel}_{exact}(k, q_2) = \frac{1}{2}\left(\mathrm{rel}_{exact}\left(k, q_2^1\right) + \mathrm{rel}_{exact}\left(k, q_2^2\right)\right) = \frac{1}{2}(0 + 1) = 0.5 \tag{13}$$

while approximate metrics provided a slightly higher score is in Equation (14):

$$\mathrm{rel}_{approx}(k, q_2) = \frac{1}{2}\left(\mathrm{rel}_{approx}\left(k, q_2^1\right) + \mathrm{rel}_{approx}\left(k, q_2^2\right)\right) = \frac{1}{2}(0.2 + 1) = 0.6 \tag{14}$$

5.2. Binomial Classification with Lift Charts

In the experiment, all queries q_i returned $\left|\dot{\mathbf{D}}\right| = 100$ images sorted in descending order according to the selected distance metric. It was essential to find the threshold value t that fixes the boundary between categories, i.e., how the retrieved pictures will be classified. Given a picture $p_i \in \dot{\mathbf{D}}$ and its rank $r_i \in [1,\ 100]$ in the retrieved results, where $r_j = 1$ indicates the most relevant picture p_j, the picture p_i was assigned to the category $cat_i = \{True, False\}$ as in Equation (15):

$$\begin{aligned} cat_i &= True,\ r_i \leq t \\ cat_i &= False,\ r_i > t \end{aligned} \tag{15}$$

Thus, in every query, the retrieved corpus was binarily classified into two subsets: (1) the category "*True*" with relevant pictures to the search query, and (2) the category "*False*" with irrelevant pictures. The threshold value (i.e., rank cutoff) was set individually for every query with a lift chart.

For the maximum precision, the cutoff was assigned to the rank with the highest lift factor. This adaptive approach assures a more objective ranking and better retrieval performance than a constant classification threshold. For example, if a result set has the maximum lift factor for rank $r = 10$ to achieve the highest precision, only samples with $r \leq 10$ should be classified as "*True*" and all others with $r > 10$ as "*False*".

This optimization approach contrasts with the fitting for recall (sensitivity), where the classification threshold is set to a relatively high value to include as many true positives as possible. In the experiment, the threshold was set at 90%. On the other hand, this approach will inevitably result in a lower precision, because many false positives will also be retrieved. It is important to point out that by using lift charts it is always possible to choose between higher precision or recall and customize the classification accordingly.

In the experiment, the classifications were optimized twice for both goals and the lift charts were split into 5% intervals. An example from the experiment is presented in Figure 5 below. Here, the dataset was queried with a single keyword "man" and ranked using the Levenshtein algorithm.

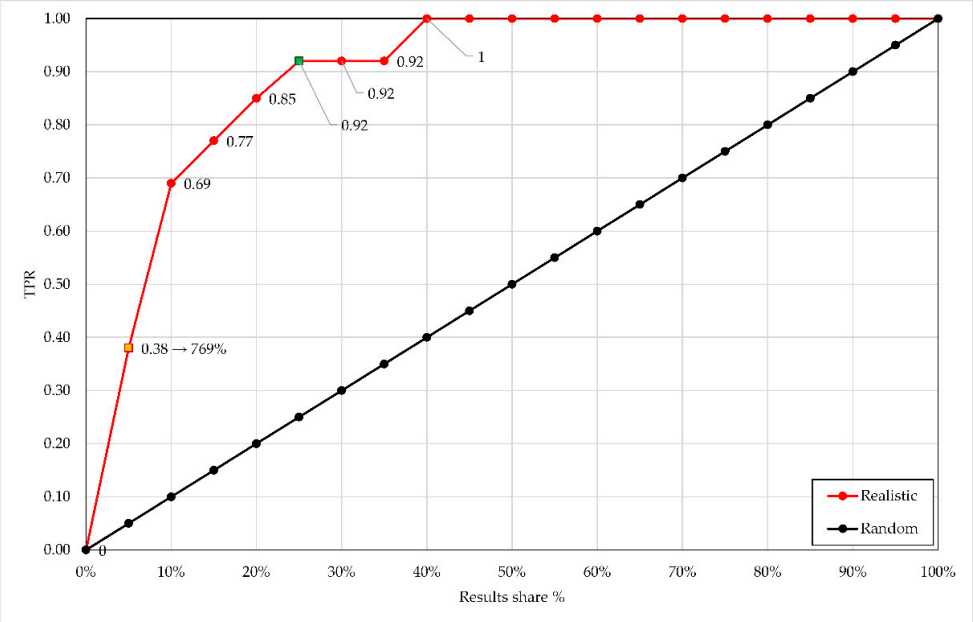

Figure 5. Lift chart of a classifier used in the study (red line) compared to random guess (black line) with the maximum lift (orange marker) showing a 769% improvement in True Positive Rate (TPR).

In Figure 5, the red curve represents the lift chart of the classifier used in the picture retrieval study and a random classifier is designated with the black line. The orange marker represents the maximum lift over a random classifier. As can be seen in Figure 5, the maximum lift of 769% was achieved for $t = 0.05$ and $TPR(t) = 0.38$. All other data segments provided lower lifts.

Additionally, in Figure 5, the green marker at 0.92 TPR represents the cutoff of optimization for recall because it indicates the lowest share of results (in this case 0.25 or 25%) where TPR ≥ 0.9. Thus, in this application for optimal recall, the cutoff must be set as $r = 25$. Again, pictures with $r \leq 25$ will be presented as search results, and all other pictures with $r > 25$ will be omitted. The recall was increased, but in a trade-off, precision and accuracy were reduced.

In the experiment, ground truth annotations of images were available, but they were semantically inadequate to fully describe the pictures and their content. Subsequently, domain experts could not agree on the true rank of images and only the assignment into the two classes was possible. This kind of problem is especially suited for lift charts, as they can be used to quantitatively evaluate and compare different classifiers, even when document rank cannot be known.

The detailed progress of the lift depending on the results share t is displayed in Figure 6. Consequently, the 5% point must be chosen for cutoff in optimization for precision: pictures with rank $r \leq 5$ should be assigned to the class "*True*" and displayed to users, and pictures with $r > 5$ to the class "*False*" and disregarded. With a cutoff at $r = 5$, $TPR = 0.38$, or in other words, only 38% of the TP pictures in the set were retrieved. This contributes to poor performance in recall but improves precision and accuracy.

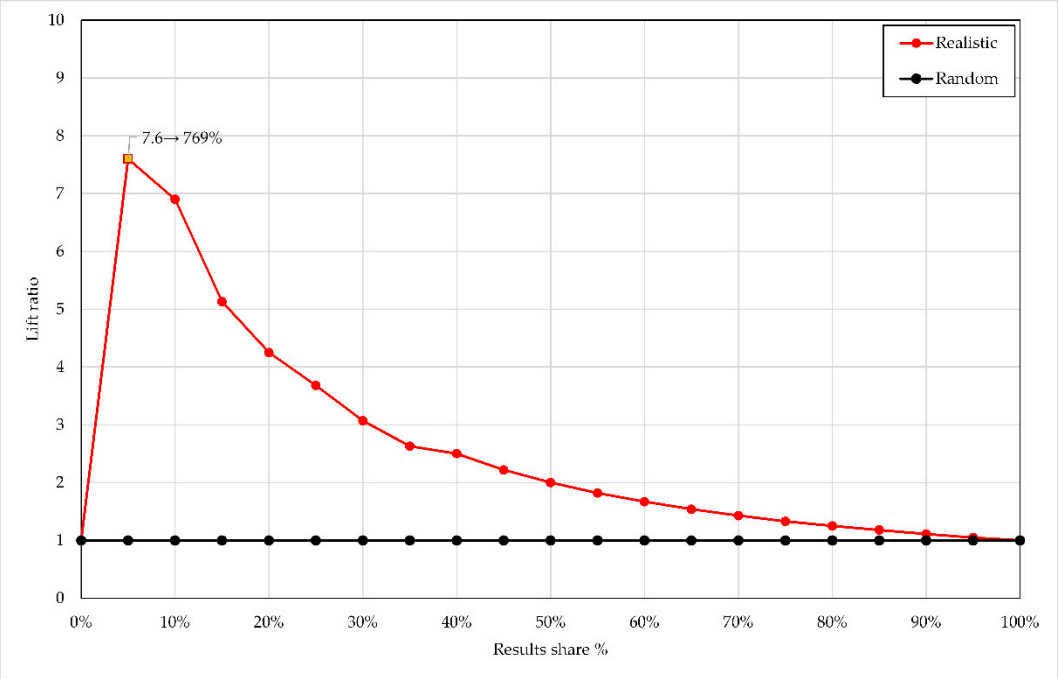

Figure 6. Lift score diagram for the lift chart of the real classifier (red series) in Figure 5. The maximum lift, as an improvement of 7.69 or 769% over random guess (black series), corresponds to the results share of 0.05 (5%).

For better clarity, the entire affective multimedia retrieval process is displayed in the UML activity diagram in Figure 7.

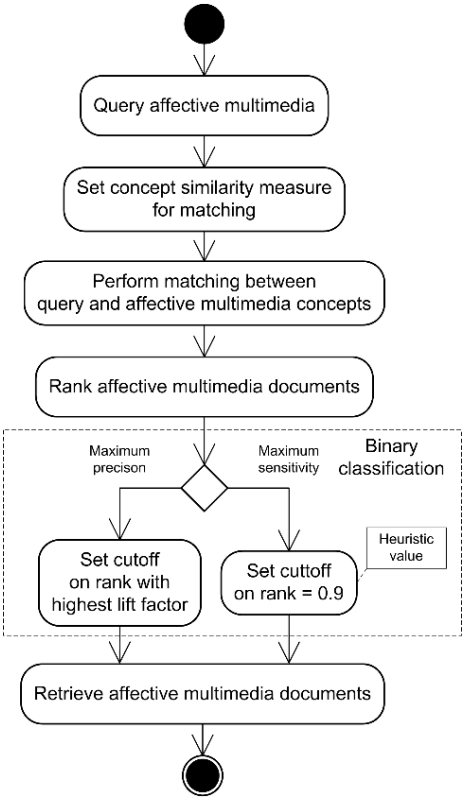

Figure 7. UML activity diagram of the affective multimedia retrieval process based on lift charts and used in the experiment.

The affective multimedia retrieval starts with a user querying the affective multimedia database using an appropriate computer interface. The users enter free-text keywords as their query and select which concept similarity measure will be used for matching. The system calculates semantic similarity between the entered query and descriptions of multimedia in the database. This step could be time-consuming, depending on the computational complexity of the chosen similarity algorithm. The duration of the ranking also linearly depends on the size of the database. After ranking has finished, affective multimedia documents are sorted in the order proportional to the similarity of their describing concepts to the concepts in the posed query. Since documents are indexed, the sorting should be less complex than the ranking. In the next step, the binary classification using lift charts is performed. If the optimization for precision is desired, then the system will find the rank with the maximum lift factor and set the cutoff at that value. Alternatively, if the search must be optimized for recall while retaining the highest achievable level of precision, then the cutoff is set at a fixed rank of 0.9. All documents with a rank less than the cutoff value are classified as positive, and all others are classified as negative. The described binary classification of affective multimedia documents is a very important step and therefore has been explained in detail in this section. Finally, in the last step, the system will fetch all positively classified documents. All documents of positively classified instances are retrieved from the multimedia repository and presented to the user. This action may also be laborious and time-consuming, but its complexity hinges on the performance of storage and data transfer subsystems and on the processor, like the ranking.

6. Experiment Results

The aggregated retrieval results optimized for recall are shown in Table 1, while those optimized for precision are shown in Table 2. Each table presents five essential performance measures (accuracy, precision, recall, fall-out, and F-measure) for two lexical similarity algorithms (exact and approximate),

which were used with one-, two-, and three-word queries. Furthermore, results from the tables are displayed as graphs in Figures 8 and 9. The experiment in [35] used a relatively similar portion of the IAPS dataset (62%) compared to this experiment (66.95%), which explains why some values obtained in these two experiments are identical. Statistically significant differences between exact and approximate lexical similarity algorithms are marked in Tables 1 and 2. In Table 3, we show the statistically significant differences between one-word, two-word, and three-word queries. Statistically insignificant differences are not depicted.

Table 1. Aggregated retrieval performance measures in classifications optimized for recall.

Query Size	Relatedness Measure	Accuracy	Precision	Recall	Fall-Out	F-Measure
1	Exact	0.2821	0.1989	0.9468	0.0444	0.3287
	Approx.	0.2775	0.2028	0.9589	0.0423	0.3348
2	Exact	0.3721	0.3364	0.9112	0.0673	0.4914
	Approx.	0.3677	0.3275	0.9569	0.0529	0.4880
3	Exact	0.4611	0.4478	0.9209	0.0761	0.6026
	Approx.	0.5106 *	0.4741 *	0.9586 *	0.0513 *	0.6344 *

* Statistically significant difference (exact vs. approximate), paired two-tailed *t*-test, $\alpha = 0.05$.

Table 2. Aggregated retrieval performance measures in classifications optimized for precision.

Query Size	Relatedness Measure	Accuracy	Precision	Recall	Fall-Out	F-Measure
1	Exact	0.7720	0.7558	0.3356	0.6661	0.4648
	Approx.	0.8183	0.6394 *	0.2668	0.7471	0.3765
2	Exact	0.7125	0.9086	0.1975	0.8189	0.3245
	Approx.	0.6911	0.8933	0.2268	0.7883	0.3617
3	Exact	0.6845	1.0000	0.2374	0.7723	0.3837
	Approx.	0.6629 *	0.9970	0.2072 *	0.8224 *	0.3430 *

* Statistically significant difference (exact vs. approximate), paired two-tailed *t*-test, $\alpha = 0.05$.

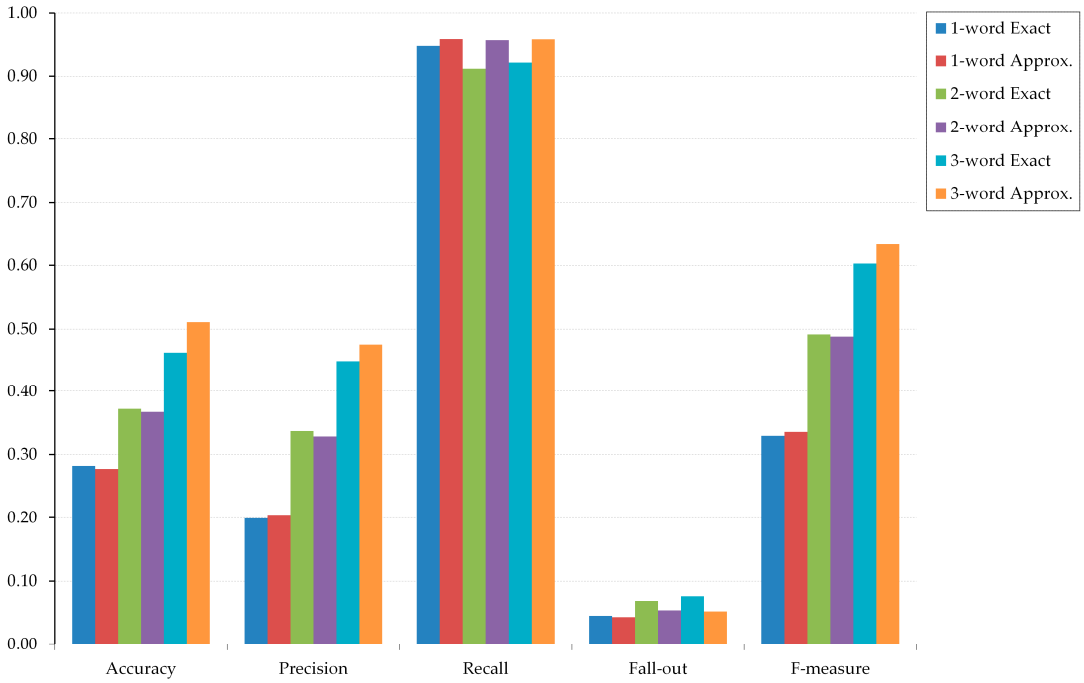

Figure 8. Comparison aggregated performance measures in retrieval optimized for recall.

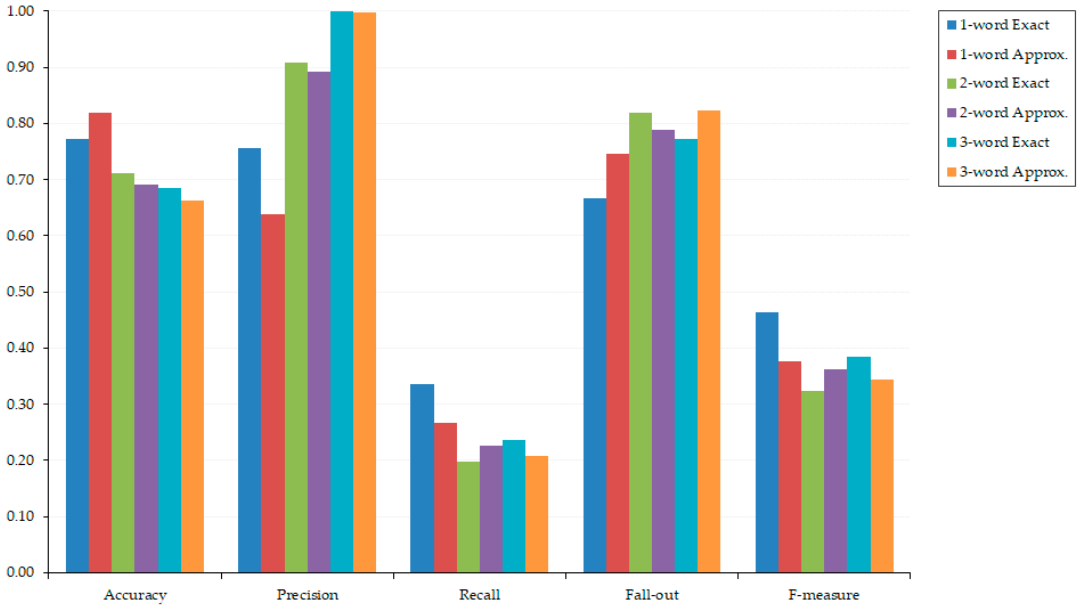

Figure 9. Comparison aggregated performance measures in retrieval optimized for precision.

Table 3. Retrieval statistical pair tests for one-way ANOVA ($\alpha = 0.05$); significant differences are marked with square symbols (▪).

Paired Sets	Significant Measures				
	Accuracy	Precision	Recall	Fall-Out	F-Measure
Optimized for Recall					
1-word vs. 2-word, Exact		▪			▪
1-word vs. 3-word, Exact	▪	▪			▪
2-word vs. 3-word, Exact	▪	▪			▪
1-word vs. 2-word, Approx.		▪			▪
1-word vs. 3-word, Approx.	▪	▪			▪
2-word vs. 3-word, Approx.	▪	▪			▪
Optimized for Precision					
1-word vs. 2-word, Exact			▪	▪	
1-word vs. 3-word, Exact	▪	▪			
2-word vs. 3-word, Exact			▪	▪	
1-word vs. 2-word, Approx.	▪	▪			
1-word vs. 3-word, Approx.	▪	▪			
2-word vs. 3-word, Approx.					

As can be seen from Table 1 and Figure 8, with classifications optimized for recall, where cutoff was set at 90% of TPR in the ranked results, multiword word queries performed better than single word queries in general: the performance of the three-word queries was better than the two-word queries, and they, in turn, were better than the queries with only one keyword. Particularly, from Table 3, we observe that there were significant improvements in accuracy, precision and F-measure between the single- and two-word queries, and between the two- and three-word queries, both for the exact and for the approximate matching algorithms (only accuracy for one- vs. two-word queries for approximate matching was insignificant). Because the classification was optimized for recall, the recall measure remained continuously high throughout all queries, with statistically insignificant variations.

As can be seen from Table 1 and Figure 8, recall was very high, while fall-out was low for all query sizes. This indicates that the retrieval algorithm is very capable of avoiding false negative outcomes. Query results systematically showed far more false positives than false negatives. This can

be explained with the prevalence of high cutoffs in optimization for recall, which entailed a small proportion of images being classified as False. Thus, false positive rate and, consequently, fall-out were almost negligible. Together, these results indicated that the retrieval optimized for recall was better in discrimination of non-relevant documents, while slightly less efficient in the identification of relevant documents.

The data in Table 1 constantly show a smaller difference in performance measures between individual matching algorithms than between single- and multi-word queries. This would suggest that the choice of a relatedness algorithm is less important to retrieval than the query size. Indeed, when optimized for recall, all measures did not show statistically significant differences in *t*-tests, except for three-word queries, where approximate matching achieved improved results. This could be explained by examining the semantic model and how approximate matching is applied to this model. Pictures are sparsely annotated with tags from an unmanaged glossary. When the query size increases, it should be expected that individual queries are more probable to approximately match with the picture tag. If the query size is smaller, this probability should also be proportionally lower. Furthermore, since the cutoff in lift charts optimized for recall does not need to be set explicitly at 90%, it is possible to change the threshold to other values (e.g., 75%, 80%, 85%, and 95%) and test how the overall retrieval performance will be affected. This seems like an interesting direction for investigation in subsequent experiments.

In the case where classification is optimized for precision—as can be seen in Table 2 and Figure 9—with a cutoff set at the rank with the highest lift factor, the approximate matching algorithms did not perform better than the naïve exact matching. Occasionally, the approximate matching results were even statistically worse than for exact matching, especially for three-word queries. This may be attributed to the fact that although the semantic model of the IAPS database contains many unique keywords and every picture is semantically described with a single keyword, a considerable proportion of pictures in the database are tagged with a common keyword. In other words, a small set of specific keywords (e.g., "man", "woman", "child") is shared among a number of pictures in the database. If, on occasion, these very keywords appear in a query with exact matching, then it could be expected that the retrieved set will show higher accuracy compared to other queries with different matching methods. Regarding query sizes, only precision showed statistically significant improvement in results between one-word and three-word queries, both for the exact and approximate matching algorithms. Recall and fall-out showed statistically significant diminished results for one-word vs. two-word queries and then improved results between two-word and three-word queries for the exact matching algorithm. Overall, the results for multiword queries were diminished with respect to single word queries for all measures except for precision, which was optimized. These results are unlike those previously reported in optimization for recall. Indeed, precision for the three-word queries in Table 2 was very high, even reaching 100% when the exact matching algorithm was used. However, this should be interpreted as an artifact of the experimental dataset, rather than a universal rule applying to all affective multimedia databases. Such very high precision results are a consequence of a very low classification threshold in almost all instances, of only 5% (i.e., the most closely related five images to the search query in $\dot{\mathbf{D}}$). In such a small sample, both algorithms could perform quite well and accurately rank documents. It can also be seen that the exact matching benefited from the choice of search keywords. Generally, in optimization for precision (in Table 2 and Figure 9), the one-word queries fared better in accuracy, recall, fall-out, and F-measure, but the three-word queries showed statistically significant better precision for both exact and approximate matching, as shown in Table 3. It could be expected that, in a realistic setting, the users of a document retrieval system would add words to queries if they would not be satisfied with the results. The results showed that, in such circumstances, the optimization for the recall algorithm will give a satisfactory performance, because adding words to query string improves recall in the text-based retrieval of affective multimedia documents.

The AUC for recall and precision depended on the shape of the lift charts. In the case of recall, the curves were much flatter in appearance, because the cutoff was set at a relatively high rank to

ensure that 90% or more TP images are returned. Conversely, lift curves adapted for precision had a much more pronounced elbow (i.e., a point with a high positive gradient), where TPR significantly increased. Such curves had a very high maximum lift. Subsequently, AUC for recall resembled more that of a random classifier than in the case of optimization for precision.

The highest attained accuracies were 51.06% and 81.83%, achieved in optimization for recall and precision, respectively. The first outcome was attained with a three-word approximate query and the second with a one-word approximate query. To improve accuracy over these values, several different actions are possible. Firstly, a semantic model could be upgraded by adding new unmanaged picture tags or by aligning tags with a managed glossary. Additionally, the existing model could be substituted with an entirely new model using linguistic networks, knowledge graphs or other knowledge representation formalisms such as ontologies. Secondly, different matching algorithms should be investigated. Correctly identifying semantic relationships between queries and image descriptors should improve accuracy in retrieval. Finally, it seems reasonable to assume that moving away from binomial logic towards reasoning with the imperfect (i.e., uncertain, imprecise, incomplete, or inconsistent) information and knowledge, which better describes the real world, might help in making a better system for retrieval of affective multimedia.

As already explained, the IAPS database may be regarded as a typical representative, or an archetype, of affective multimedia databases. Semantic and emotional image annotation models of almost all other such databases are virtually identical to IAPS. For these reasons, image retrieval performance achieved with IAPS may be considered indicative for all comparable datasets developed for experimentation in emotion and attention research.

In summary, the overall results suggest that tagging pictures with only one keyword from unsupervised glossaries gives poor information retrieval performance regardless of the classifier optimization. In such a sparse labeling approach, false positives and false negatives may be frequent. Moreover, query expansion from 1 to 3 keywords per query was shown to univocally improve precision, while better accuracy was possible only in classification optimized for recall. The recommended strategy for the retrieval of affective pictures may be summed up in the following two rules. First, the optimization for precision is the default type of retrieval if lift charts are used in the unsupervised mode for concept-based image retrieval. Second, if higher accuracy is important, then the query must be expanded with additional keywords and the classification optimized for recall should be the preferred choice.

7. Conclusions

Using a real-life problem situation of concept-based retrieval of emotionally-annotated images from the IAPS database, we demonstrated that lift charts are helpful tools in the binary classification of affective pictures, i.e., text-based retrieval of pictures from affective multimedia databases. Lift charts are algorithmically inexpensive and can be universally applied to a range of concept-based and text-based retrieval tasks. The classification can be adapted to achieve better accuracy or higher retrieval if a larger proportion of true positive results or a higher number of all items is needed, respectively. Apart from commonplace retrieval quality indicators, such as accuracy, precision, recall, fall-out, F-measure, mean average precision, and discounted cumulative gain, individual classifiers may be evaluated and mutually compared with lift charts using maximum lift factor and AUC measures, thus providing additional useful information in the selection of optimal classifiers for specific retrieval tasks. Newly developed databases, listed in Section 2, still suffer from an informal and weakly expressive representation model. Retrieval methods implemented in these databases must cope with these defects. Nonetheless, our experiment has shown that lift charts are a beneficial technique for evaluating and improving retrieval performance even in such an unfavorable setting.

We have also shown that, although lift charts are very successful in assessment of classifiers' performance, they cannot always be considered optimal in the rank retrieval of emotionally-annotated images. In this regard, their performance highly depends on the context in which they are used, i.e., external factors related to the characteristics and functioning of a retrieval system. For example,

lift charts as classifiers cannot overcome problems that may arise from a vague or inadequate representation of knowledge about the indexed documents and their content.

To conclude, lift charts are very helpful in the evaluation of classifiers in circumstances where true categories of documents cannot be known. However, if they are used as classifiers, more caution is necessary, and their real performance should be objectively analyzed post-hoc.

Author Contributions: Conceptualization, M.H.; methodology, M.H. and A.J.; software, M.H.; validation, M.H., A.J. and D.I.; formal analysis, M.H. and A.J.; investigation, M.H., A.J. and D.I.; resources, M.H.; data curation, M.H.; writing—original draft preparation, M.H.; writing—review and editing, M.H., A.J. and D.I.; visualization, M.H. and A.J.; supervision, M.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Manning, C.D.; Raghavan, P.; Schütze, H. *Introduction to Information Retrieval*; Cambridge University Press: Cambridge, UK, 2008.
2. Munir, K.; Anjum, M.S. The use of ontologies for effective knowledge modelling and information retrieval. *Appl. Comput. Inform.* **2018**, *14*, 116–126. [CrossRef]
3. Vuk, M.; Curk, T. ROC curve, lift chart and calibration plot. *Metodoloski Zvezki* **2006**, *3*, 89.
4. Microsoft, Lift Chart (Analysis Services—Data Mining). Available online: https://docs.microsoft.com/en-us/sql/analysis-services/data-mining/lift-chart-analysis-services-data-mining (accessed on 1 June 2020).
5. Witten, I.H.; Frank, E. Data mining: Practical machine learning tools and techniques with Java implementations. *ACM Sigmod Rec.* **2002**, *31*, 76–77. [CrossRef]
6. Mezaris, V.; Kompatsiaris, I.; Strintzis, M.G. An ontology approach to object-based image retrieval. In Proceedings of the 2003 International Conference on Image Processing, Barcelona, Spain, 14–17 September 2003; Volume 2, pp. II–511.
7. Datta, R.; Li, J.; Wang, J.Z. Content-based image retrieval: Approaches and trends of the new age. In Proceedings of the 7th ACM SIGMM International Workshop on Multimedia Information Retrieval, Singapore, 10–11 November 2005; pp. 253–262.
8. Horvat, M. A Brief Overview of Affective Multimedia Databases. In *Central European Conference on Information and Intelligent Systems*; Faculty of Organization and Informatics: Varaždin, Croatia, 2017; pp. 3–9.
9. Lang, P.J.; Bradley, M.M.; Cuthbert, B.N. International affective picture system (IAPS): Technical manual and affective ratings. *NIMH Cent. Study Emot. Atten.* **1997**, *1*, 39–58.
10. Lang, P.; Bradley, M.M. The International Affective Picture System (IAPS) in the study of emotion and attention. In *Series in Affective Science. Handbook of Emotion Elicitation and Assessment*; Coan, J.A., Allen, J.J.B., Eds.; Oxford University Press: New York, NY, USA, 2007; pp. 29–46.
11. Colden, A.; Bruder, M.; Manstead, A.S. Human content in affect-inducing stimuli: A secondary analysis of the international affective picture system. *Motiv. Emot.* **2008**, *32*, 260–269. [CrossRef]
12. Marchewka, A.; Żurawski, Ł.; Jednorog, K.; Grabowska, A. The Nencki Affective Picture System (NAPS): Introduction to a novel, standardized, wide-range, high-quality, realistic picture database. *Behav. Res. Methods* **2014**, *46*, 596–610. [CrossRef]
13. Riegel, M.; Żurawski, Ł.; Wierzba, M.; Moslehi, A.; Klocek, Ł.; Horvat, M.; Grabowska, A.; Michałowski, J.; Marchewka, A. Characterization of the Nencki Affective Picture System by discrete emotional categories (NAPS BE). *Behav. Res. Methods* **2016**, *48*, 600–612. [CrossRef]
14. Wierzba, M.; Riegel, M.; Pucz, A.; Leśniewska, Z.; Dragan, W.Ł.; Gola, M.; Jednorog, K.; Marchewka, A. Erotic subset for the Nencki Affective Picture System (NAPS ERO): Cross-sexual comparison study. *Front. Psychol.* **2015**, *6*, 1336. [CrossRef]
15. Dan-Glauser, E.S.; Scherer, K.R. The Geneva affective picture database (GAPED): A new 730-picture database focusing on valence and normative significance. *Behav. Res. Methods* **2011**, *43*, 468–477. [CrossRef]
16. Miccoli, L.; Delgado, R.; Guerra, P.; Versace, F.; Rodríguez-Ruiz, S.; Fernández-Santaella, M.C. Affective pictures and the Open Library of Affective Foods (OLAF): Tools to investigate emotions toward food in adults. *PLoS ONE* **2016**, *11*, e0158991. [CrossRef]

17. Haberkamp, A.; Glombiewski, J.A.; Schmidt, F.; Barke, A. The DIsgust-RelaTed-Images (DIRTI) database: Validation of a novel standardized set of disgust pictures. *Behav. Res. Ther.* **2017**, *89*, 86–94. [CrossRef] [PubMed]
18. Michałowski, J.M.; Droździel, D.; Matuszewski, J.; Koziejowski, W.; Jednorog, K.; Marchewka, A. The Set of Fear Inducing Pictures (SFIP): Development and validation in fearful and nonfearful individuals. *Behav. Res. Methods* **2017**, *49*, 1407–1419. [CrossRef] [PubMed]
19. Kurdi, B.; Lozano, S.; Banaji, M.R. Introducing the open affective standardized image set (OASIS). *Behav. Res. Methods* **2017**, *49*, 457–470. [CrossRef]
20. Zamora, E.V.; Richard's, M.M.; Introzzi, I.; Aydmune, Y.; Urquijo, S.; Olmos, J.G.; Marchewka, A. The Nencki Affective Picture System (NAPS): A Children-Rated Subset. *Trends Psychol.* **2020**, 1–17. [CrossRef]
21. Mehrabian, A. Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament. *Curr. Psychol.* **1996**, *14*, 261–292. [CrossRef]
22. Peter, C.; Herbon, A. Emotion representation and physiology assignments in digital systems. *Interact. Comput.* **2006**, *18*, 139–170. [CrossRef]
23. Bakker, I.; van der Voordt, T.; Vink, P.; de Boon, J. Pleasure, arousal, dominance: Mehrabian and Russell revisited. *Curr. Psychol.* **2014**, *33*, 405–421. [CrossRef]
24. Burkhardt, F.; Pelachaud, C.; Schuller, B.W.; Zovato, E. EmotionML. In *Multimodal Interaction with W3C Standards*; Springer: Cham, Germany, 2017; pp. 65–80.
25. Long, F.; Zhang, H.; Feng, D.D. Fundamentals of content-based image retrieval. In *Multimedia Information Retrieval and Management*; Springer: Berlin/Heidelberg, Germany, 2003; pp. 1–26.
26. Bergmann, R.; Gil, Y. Similarity assessment and efficient retrieval of semantic workflows. *Inf. Syst.* **2014**, *40*, 115–127. [CrossRef]
27. Ceri, S.; Bozzon, A.; Brambilla, M.; Della Valle, E.; Fraternali, P.; Quarteroni, S. *Web Information Retrieval*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2013.
28. Pillai, S.G.; Soon, L.K.; Haw, S.C. Comparing DBpedia, Wikidata, and YAGO for Web Information Retrieval. In *Intelligent and Interactive Computing*; Springer: Singapore, 2019; pp. 525–535.
29. Fawcett, T. An introduction to ROC analysis. *Pattern Recognit. Lett.* **2006**, *27*, 861–874. [CrossRef]
30. Devriendt, F.; Berrevoets, J.; Verbeke, W. Why you should stop predicting customer churn and start using uplift models. *Inf. Sci.* **2019**. [CrossRef]
31. Yeh, I.C.; Lien, C.H. The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert Syst. Appl.* **2009**, *36*, 2473–2480. [CrossRef]
32. Boom, B.J.; Huang, P.X.; He, J.; Fisher, R.B. Supporting ground-truth annotation of image datasets using clustering. In Proceedings of the 21st International Conference on Pattern Recognition, Tsukuba, Japan, 11–15 November 2012; pp. 1542–1545.
33. Wang, C.; Liu, P.S. Data Mining and Hotspot Detection in an Urban Development Project. *J. Data Sci.* **2008**, *6*, 389–414.
34. Hakak, S.I.; Kamsin, A.; Shivakumara, P.; Gilkar, G.A.; Khan, W.Z.; Imran, M. Exact String Matching Algorithms: Survey, Issues, and Future Research Directions. *IEEE Access* **2019**, *7*, 69614–69637. [CrossRef]
35. Horvat, M.; Vuković, M.; Car, Ž. Evaluation of keyword search in affective multimedia databases. In *Transactions on Computational Collective Intelligence XXI*; Springer: Berlin/Heidelberg, Germany, 2016; pp. 50–68.
36. Feng, D.; Siu, W.C.; Zhang, H.J. *Multimedia Information Retrieval and Management: Technological Fundamentals and Applications*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2013.