

Investigating non-classical correlations between decision fused multi-modal documents.

Dimitris Gkoumas¹, Sagar Uprety¹, and Dawei Song^{1,2}

¹ The Open University, Milton Keynes, UK

{dimitris.gkoumas, sagar.uprety, dawei.song}@open.ac.uk

² Beijing Institute of Technology, Beijing, China

Abstract. Correlation has been widely used to facilitate various information retrieval methods such as query expansion, relevance feedback, document clustering, and multi-modal fusion. Especially, correlation and independence are important issues when fusing different modalities that influence a multi-modal information retrieval process. The basic idea of correlation is that an observable can help predict or enhance another observable. In quantum mechanics, quantum correlation, called entanglement, is a sort of correlation between the observables measured in atomic-size particles when these particles are not necessarily collected in ensembles. In this paper, we examine a multimodal fusion scenario that might be similar to that encountered in physics by firstly measuring two observables (i.e., text-based relevance and image-based relevance) of a multi-modal document without counting on an ensemble of multi-modal documents already labeled in terms of these two variables. Then, we investigate the existence of non-classical correlations between pairs of multi-modal documents. Despite there are some basic differences between entanglement and classical correlation encountered in the macroscopic world, we investigate the existence of this kind of non-classical correlation through the Bell inequality violation. Here, we experimentally test several novel association methods in a small-scale experiment. However, in the current experiment we did not find any violation of the Bell inequality. Finally, we present a series of interesting discussions, which may provide theoretical and empirical insights and inspirations for future development of this direction.

Keywords: Multi-modal information retrieval · Non-classical correlations · Decision fused multi-modal documents · CHSH inequality

1 Introduction

Nowadays, the Web surrounding us often involves multiple modalities - we read texts, watch images and videos, and listen to sounds. In general terms, modality refers to a certain type of information and/or the representation format in which information is stored. A research problem is characterized as multi-modal when it includes multiple such modalities. Integrating unimodal representations from various input modalities and combining them into a compact multi-modal

representation, called multi-modal fusion, offers a possibility of understanding in-depth real world problems. For instance, in information retrieval, suppose a user types in a text query to retrieve multi-modal documents consisting of an image and a caption as shown in Fig. 1. One can notice that the query term “plane” can be matched in both textual and visual modalities of the given multi-modal document. However, the query term “LHR” can be matched only in its textual modality, while the term “sunset” only in its visual modality. This implies that only when the text and image modalities are fused, we get the benefit of complementary information, in turn increasing the precision of information retrieval.

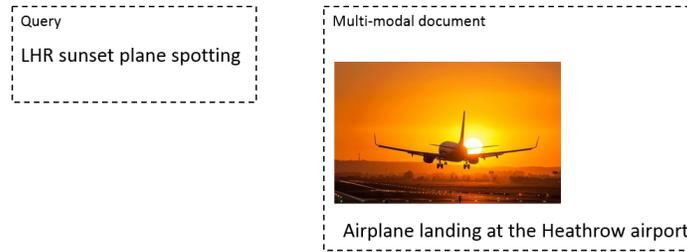


Fig. 1. Example of multi-modal information retrieval

The main challenge of multi-modal fusion is to capture inter-dependencies and complementary presence in heterogeneous data originating from multiple modalities. In the literature, two main approaches to the fusion process have been proposed: a) *feature level* or *early fusion* and b) *decision level* or *late fusion* [4]. Early fusion involves the integration of multiple sources of raw or preprocessed data to be fed into a model, which finally makes an inference as illustrated in Fig. 2. In contrast, late fusion refers to the aggregation of decisions from multiple classifiers, each trained on separate modalities as shown in Fig. 3.

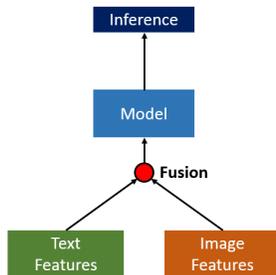


Fig. 2. Early fusion

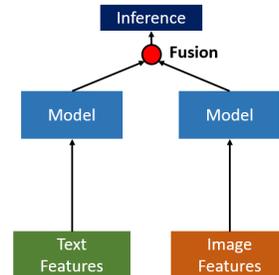


Fig. 3. Late fusion

There are distinctive issues that influence the multi-modal fusion process. Correlation between different modalities is one of them. Correlation can be perceived either in low-level features, e.g., raw data, or high-level features that are obtained on different classifiers, e.g., semantic concepts [4]. In both cases, correlation informs us how to fuse different modalities. In the early fusion, we fuse multi-modal information either by projecting all of the modalities to the same space (Fig. 4 (c)), called joint representations, or by learning separate representations for each modality but coordinate them through a similarity measure (Fig 4 (b)) [5]. In both approaches, the construction of the multi-modal spaces is based on correlations among different modalities. The late fusion process can be rule-based, e.g., by linear weighted fusion and majority voting rules, or based on classification-based methods, e.g., support vector machines [4]. In many cases, the correlation among different modalities provides additional cues that are very useful for aggregating decisions either by following a rule-based approach or classification-based approach. In addition, the absence of correlation may equally provide valuable insight with respect to a particular scenario or context.

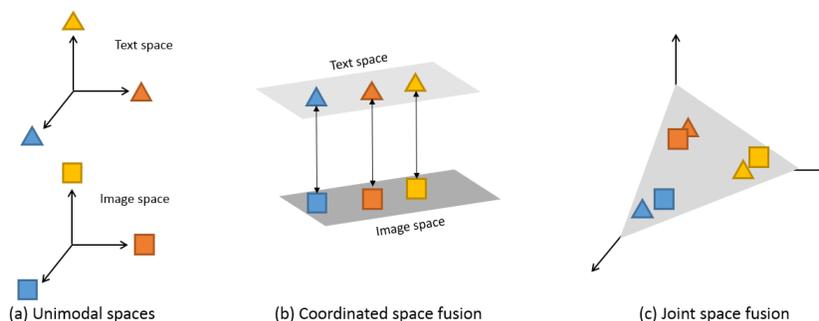


Fig. 4. Construction of multi-modal spaces

There are various statistical and probabilistic forms of correlation that have been utilized by researchers, being causal or not. Since our experiment focuses on late fusion only, we briefly report the most important methods for computing correlations between decisions from multiple modalities. Specifically, decision level correlation has been exploited in the form of causal link analysis, causal strength, and agreement coefficient [4]. In all cases, the basic idea of correlation is that a modality can help predict or enhance another modality.

In quantum mechanics, correlation has been also an important topic. In quantum mechanical framework, uncertainty may occur not only when the elements are collected in ensemble but also when each of them is in a superposed state. In quantum theory, making an observation on one part of a system *instantaneously* could affect the state in another part of a system, even if the respective systems are separated by space-like distances. Such a quantum correlation presents some peculiarities which led to the notion of entanglement. Entanglement is a

sort of correlation between observables measured in atomic-size particles, such as photons, when these particles are not necessarily collected in ensembles.

Despite entanglement being a kind of correlation, there are some basic differences between entanglement and the classical correlation encountered in the macroscopic world. A classical correlation is a statistical relationship, causal or not, between two random variables. In entanglement, besides correlation, cause exists as well since the correlation does not depend on an underlying value attached to the particles. Instead, it depends on what is measured on either side. This non-classical property of quantum entanglement motivates us to investigate non-classical correlations between multi-modal decisions as shown in Fig. 5. At first, we calculate the probability of relevance for each document, with respect to both text-based and image-based modality concerning a multimodal query as shown in Fig. 5. Then, we check for any violation of Bell’s inequalities based on the estimated relevance probabilities for each possible pair of decision fused multimodal documents in a dataset. Our assumption is that if a pair of decision fused multi-modal documents is *entangled*, then knowing that a document is relevant concerning the text-based representation for a query, then we can *simultaneously* predict with certainty the relevance of the the other document concerning the text-based and image-based representation for the same query.

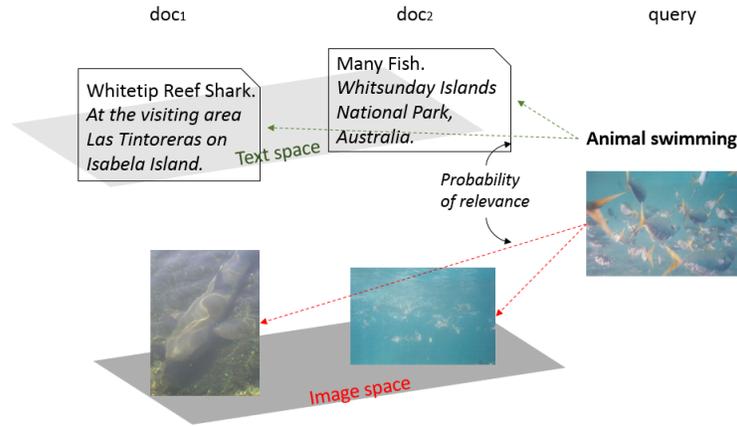


Fig. 5. Investigation of non-classical correlations between decision fused multi-modal documents

The rest of the paper is organized as follows. Section 2 presents a brief review of related work. In Section 3 we provide a foundation in quantum entanglement and Bell inequality, while in Section 4 we explain some basic concepts of geometry in information retrieval and then formalize the proposed model. Section 5 reports all the experiment settings. In Section 6 we report and discuss the results. Finally, Section 7 concludes the paper.

2 Related Work

A composite system being entangled cannot be validly decomposed and modeled as separate subsystems. The quantum theory provides formal tools to model interacting systems as non-decomposable in macroscopic world as well. The phenomenon of quantum entanglement has been investigated in semantic spaces making use of Hyperspace Analogue to Language (HAL) model [13,14]. Hou et al. considered high order entanglements that cannot be reduced to the compositional effect of lower-order ones, as an indicator of high-level semantic entities. Melucci proposes quantum-like entanglement for modeling the interaction between a user and a document as a composite system [15].

The non-compositionality of entangled systems opened also the door to developing quantum-like models of cognitive phenomena which are not decompositional in nature. Concept combinations have been widely modeled as composite systems [1,2,6,7,22]. The state of the composite system between two words can be modeled by taking the tensor product of the states of the individual words respectively. If the concept combination is factorizable, then the concept is compositional in the sense it can be expressed as a product of states corresponding to the separate words. A concept that is not factorizable cannot be expressed by either the first or the second word individually, and is deemed *non-compositional*, and termed *entangled* [7].

Quantum theory provides a well-developed set of analytical tools that can be used to determine whether the state of a system of interest can be validly decomposed into separate sub-systems. A possible way to test the non-compositional state of a composite system is the violation of Bell's inequalities. For instance, having calculated the expectation values of variables associated with an experiment, we can fit the Clauser-Horne-Shimony-Holt (CHSH) version of Bell's inequality [9]. If the CHSH inequality is greater than 2, then the Bell inequality is violated. It has been empirically found that the maximal possible violation in quantum theory is $2\sqrt{2} \approx 2.8284$ [8]. This means that each violation being close to the maximal value is very significant. In addition to the CHSH inequality, Bruza et al. [7] propose Clauser-Horne inequalities to analyse the decomposability of quantum systems. The Schmidt decomposition is another way for detecting entanglement in bipartite systems [17]. According to the theorem, after decomposition, each pure state of the tensor product space can be expressed as the product of subsystem orthonormal bases and non-negative real coefficients. The square sum of the coefficients is equal to 1. The number of non-zero coefficients is called Schmidt number. If it equals 1, then the composite state is the product state. If it greater than 1, then the composite state is non-compositional.

So far, researchers have used joint probabilities in cognitive science for calculating expectation values assuming that the outcomes of observables are dependent. Additionally, probabilities can be calculated via trace formula in Gleason's theory [11]. In a similar way, expectation value of two random variables is defined the product of traces [15]. Finally, probabilities could be re-expressed as function of an angle θ , where θ is defined as a difference in phase between two

random observables, once we view the relationship between them as a geometrical relationship [15].

3 Quantum Entanglement and Bell Inequality

Let us suppose that we have a system of two qubits expressed in a Bit basis $\{0, 1\}$, such that the first qubit is in a state $a_0|0\rangle + a_1|1\rangle$ and the second one in a state $b_0|0\rangle + b_1|1\rangle$. The state of the two qubits together as a composite system is a superposition of four classical probabilities resulting in

$$|\phi\rangle = a_0b_0|00\rangle + a_0b_1|01\rangle + a_1b_0|10\rangle + a_1b_1|11\rangle. \quad (1)$$

Let us now assume that the composite system is in an entangled state given by the following Bell state

$$|\psi\rangle = \frac{1}{\sqrt{2}}|00\rangle + \frac{1}{\sqrt{2}}|11\rangle. \quad (2)$$

When we measure the composite system, the probability of the system to collapse either to the state $|00\rangle$ or to the state $|11\rangle$ is equal to 0.5. However, after a measurement, the system is not in an entangled state anymore. For instance, once we measure the state $|00\rangle$, the new state of the system results in

$$|\psi\rangle = |00\rangle. \quad (3)$$

Let us now assume that we measure the state $|0\rangle$ of the first qubit (equation (2)). Then the probability for the first qubit to collapse to the state $|0\rangle$ again equals 0.5. However, after the measurement, the probability of the second qubit to be in the state $|0\rangle$ currently equals 1. Let us suppose that we change the Bit basis to a Sign basis $\{-, +\}$. According to the rotation invariance [19], the Bell state in the Sign basis is again an equal superposition of the state $|--\rangle$ and the state $|++\rangle$ such that

$$\begin{aligned} |\psi\rangle &= \frac{1}{\sqrt{2}}|00\rangle + \frac{1}{\sqrt{2}}|11\rangle \\ &= \frac{1}{\sqrt{2}}|--\rangle + \frac{1}{\sqrt{2}}|++\rangle. \end{aligned} \quad (4)$$

Suppose now that we want to measure the probability of the second qubit to be in the state $|-\rangle$ according to the Sign basis, given that we have already measured the probability of the first qubit to be in state $|0\rangle$ concerning the Bit basis. Once we measure the first qubit, the probability of the second qubit to be in the same state $|0\rangle$ is equal to 1. If θ is the angle between the Bit and Sign basis, then according to the Pythagorean theorem, the probability of the second qubit to be in the state $|-\rangle$ equals $\cos^2 \theta$.

In quantum mechanics, the criteria used to test entanglement are given by Bell's inequalities. A possible way to proceed is to define four observables. Each observable has binary values ± 1 thus give two mutually exclusive outcomes. For

instance, a photon can be detected by ‘+’ or ‘-’ channel (see Fig. 9). Let us denote as A_1 , B_1 the observables describing the first system, and A_2 , B_2 the observables of the second one. If a composite system is separable, the following CHSH inequality holds:

$$|\langle A_1 A_2 \rangle + \langle A_2 B_1 \rangle + \langle A_1 B_2 \rangle - \langle B_1 B_2 \rangle| \leq 2, \quad (5)$$

where $\langle \rangle$ denotes the expectation value between two observables. The calculation of expectation values will be articulated in Section 4. The violation of (5) is a sign of entanglement. A Bell inequality violation implies that at least one of the assumptions of *local-realism* made in the proof of (5) must be incorrect [16]. This points to the conclusion that either or both of locality - an object is only directly influenced by its immediate surroundings - and realism - an object has definite values - must be rejected as a property of the composite systems violating CHSH inequality.

4 Non-classical Correlations in Decision Fused Multimodal Documents

Before the late fusion process, there exists a probability $p(R|T)$ for a multimodal document D_M to be relevant to a multimodal information need concerning the textual information. Similarly, the probability for the same document not to be relevant is denoted as $p(\bar{R}|T)$, which is equal to $1 - p(R|T)$. Let us consider a real-valued two dimensional Hilbert Space for the relevance of the D_M concerning the textual information (Fig. 6). In Fig. 6 the vector R_t stands for the relevance of the document concerning the text-based modality. On the other hand, the \bar{R}_t represents the non-relevance with respect to the same text-based information need and is orthogonal to R_t .

The text-based relevance of a document can be modeled as a vector in the Hilbert Space, which unifies the logical, probabilistic and vector space based approaches to IR [21]. This vector is a superposition of relevance and non-relevance vectors with respect to the text-based modality and is represented as:

$$|D_M\rangle = a|R_t\rangle + a'|\bar{R}_t\rangle, \quad (6)$$

where $|a|^2 + |a'|^2 = 1$. The coefficients a and a' are captured by taking the projection of $|D_M\rangle$ onto the relevance and non-relevance vectors respectively (Fig. 6) by taking their inner products. According to the Born rule, $p(R|T)$ equals to the square of the inner product $|\langle R_t|D_M\rangle|^2$ and likewise, $p(\bar{R}|T)$ equals to $|\langle \bar{R}_t|D_M\rangle|^2$.

In a similar way, we denote as $p(R|I)$ the probability for a multimodal document D_M to be relevant concerning the image-based information need, and $p(\bar{R}|I)$ the probability to be irrelevant respectively (Fig. 7). The relevance of a document with respect to the image-based modality can be in a similar manner modeled as:

$$|D_M\rangle = b|R_i\rangle + b'|\bar{R}_i\rangle \quad (7)$$

In this case, $p(R|I)$ is computed as the square of the inner product $|\langle R_i|D_M\rangle|^2$. Likewise, $p(\bar{R}|I)$ equals to $|\langle \bar{R}_i|D_M\rangle|^2$.

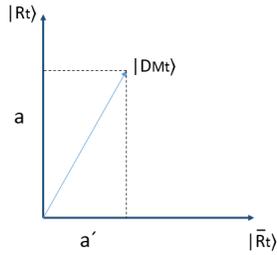


Fig. 6. Text-based relevance in two-dimensional Hilbert Space

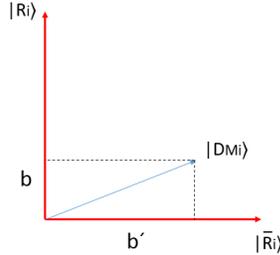


Fig. 7. Image-based relevance in two-dimensional Hilbert Space

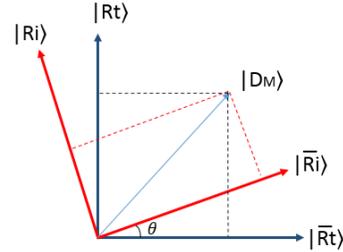


Fig. 8. Hilbert Space after fusion having a text and image basis

After the late fusion process, the document can be judged based on both text-based and image-based modalities. Such a phenomenon can be modeled in the same Hilbert Space by having a different basis for each modality, as presented in Fig. 8. The document D_M is represented as a unit vector and its representation is expressed with respect to the bases $T = \{|R_t\rangle, |\bar{R}_t\rangle\}$ and $I = \{|R_i\rangle, |\bar{R}_i\rangle\}$ fusing at the end the local decisions. Each basis models context with respect to a given modality.

The rest of the experimental setup is analogous to that one for investigating quantum entanglement in photons [3]. Fig. 9 shows the experimental setup for the violation of Bell's inequalities. The source S produces a pair of photons, sent in opposite directions. Each photon encounters a two-channel polariser whose orientation can be set by the experimenter. Coincidences (simultaneous detections) are recorded, the results being categorised as $++$, $+-$, $-+$, or $--$ and corresponding counts accumulated by the coincidence monitor CM.

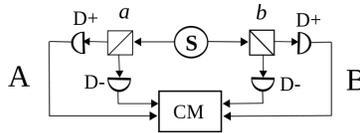


Fig. 9. Schematic of a “two-channel” Bell test

Now let us consider Fig. 10, which depicts two multimodal documents, D_{M1} and D_{M2} respectively. As afore-mentioned, the documents D_{M1} and D_{M2} can be expressed with either the text-based basis or image-based basis being in a superposition of relevance and non-relevance states. In quantum theory, the

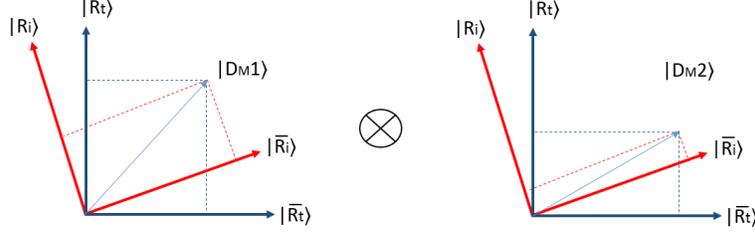


Fig. 10. The interaction between two documents is modeled as a composite system

interaction between D_{M1} and D_{M2} can be modeled as a composited system by using the tensor product of the document Hilbert Spaces. The state of the composite system $D_{M1} \otimes D_{M2}$ can be obtained by taking the tensor product of the relevance and non-relevance states. Concerning the text-based modality, the state of the composite system is defined as follows:

$$\begin{aligned} |D_{M1}\rangle \otimes |D_{M2}\rangle &= (a_1|R_t\rangle + a'_1|\bar{R}_t\rangle) \otimes (a_2|R_t\rangle + a'_2|\bar{R}_t\rangle) \\ &= a_1a_2|R_tR_t\rangle + a_1a'_2|R_t\bar{R}_t\rangle + a'_1a_2|\bar{R}_tR_t\rangle + a'_1a'_2|\bar{R}_t\bar{R}_t\rangle, \end{aligned} \quad (8)$$

where $|a_1a_2|^2 + |a_1a'_2|^2 + |a'_1a_2|^2 + |a'_1a'_2|^2 = 1$. In a similar way, if we define the image-based basis as a standard basis then the state of the composite system $D_{M1} \otimes D_{M2}$ concerning the image-based modality can be expressed as follows:

$$\begin{aligned} |D_{M1}\rangle \otimes |D_{M2}\rangle &= (b_1|R_i\rangle + b'_1|\bar{R}_i\rangle) \otimes (b_2|R_i\rangle + b'_2|\bar{R}_i\rangle) \\ &= b_1b_2|R_iR_i\rangle + b_1b'_2|R_i\bar{R}_i\rangle + b'_1b_2|\bar{R}_iR_i\rangle + b'_1b'_2|\bar{R}_i\bar{R}_i\rangle \end{aligned} \quad (9)$$

In Equation (8), the first and second terms reveal that when the text-based content of the D_{M1} is relevant, then we cannot be sure about the relevance of the text-based content of the other document. Similarly, according to the third and fourth term in Equation (8), when the text-based content of the D_{M1} is non-relevant, then the other document is in a superposition of relevance and non-relevance states with respect to the text-based modality. The same is observed when we consider the image-based basis as a standard basis.

If the state of the text-based (Equation (8)) or image-based (Equation (9)) composite system is factorizable, then the system is compositional in the sense it can be expressed as a product of states corresponding to the separate subsystems. A composite quantum system that is not factorizable is deemed *non-compositional* and termed *entangled* [7]. In the last case, if we consider the text-based representation as a standard basis, then we can define two Bell states, either the state

$$|D_M\rangle = a_1a_2|R_tR_t\rangle + a'_1a'_2|\bar{R}_t\bar{R}_t\rangle, \quad (10)$$

or

$$|D_M\rangle = a_1a'_2|R_t\bar{R}_t\rangle + a'_1a_2|\bar{R}_tR_t\rangle. \quad (11)$$

Concerning the Equation (10), the probability for both documents to be relevant (i.e., the state $|R_t R_t\rangle$) regarding the text-based modality equals $|a_1 a_2|^2$. If we measure only the probability of the first document to be relevant concerning the text-based modality results again in $|a_1 a_2|^2$. Then after the measurement, the probability for the second document to be relevant is equal to 1. Consequently, we can *simultaneously* predict the probability of the second document to be relevant concerning the image-based modality, which is equal to $\cos^2 \theta$, where θ is the angle between the image-based and text-based basis (Fig. 10). Similar outcomes result once we measure the probability for both documents to be irrelevant (i.e., the state $|\overline{R_t R_t}\rangle$ in Equation (10)), one relevant and the other irrelevant (i.e., the state $|R_t \overline{R_t}\rangle$ in Equation (11)), or one irrelevant and the other relevant (i.e., the state $|\overline{R_t R_t}\rangle$ in Equation (11)).

In Section 3, we have described the CHSH inequality defining four observables, where each observable has two binary values ± 1 thus gives two mutually exclusive outcomes. In a similar manner, in our case, for the document D_{M1} , we have variables R_{t1} and R_{i1} , which take values 1,-1, where $R_{t1} = 1$ corresponds to the basis vector $|R_{t1}\rangle$ and $R_{t1} = -1$ corresponds to its orthogonal basis vector $|\overline{R_{t1}}\rangle$. Similarly, $R_{i1} = 1$ corresponds to the basis vector $|R_{i1}\rangle$ and $R_{i1} = -1$ corresponds to its orthogonal basis vector $|\overline{R_{i1}}\rangle$. For the document D_{M2} , we have variables R_{t2} and R_{i2} which take values 1,-1, where $R_{t2} = 1$ corresponds to the basis vector $|R_{t2}\rangle$ and $R_{t2} = -1$ corresponds to its orthogonal basis vector $|\overline{R_{t2}}\rangle$. Similarly, $R_{i2} = 1$ corresponds to the basis vector $|R_{i2}\rangle$ and $R_{i2} = -1$ corresponds to its orthogonal basis vector $|\overline{R_{i2}}\rangle$. Then Equation (5) results in

$$|\langle R_{t1} R_{t2} \rangle + \langle R_{t2} R_{i1} \rangle + \langle R_{t1} R_{i2} \rangle - \langle R_{i1} R_{i2} \rangle| \leq 2, \quad (12)$$

where

$$\begin{aligned} \langle R_{t1} R_{t2} \rangle &= ((+1)p(R_{t1}) + (-1)p(\overline{R_{t1}})) * ((+1)p(R_{t2}) + (-1)p(\overline{R_{t2}})) \\ &= p(R_{t1})p(R_{t2}) - p(R_{t1})p(\overline{R_{t2}}) - p(\overline{R_{t1}})p(R_{t2}) + p(\overline{R_{t1}})p(\overline{R_{t2}}), \end{aligned}$$

$$\begin{aligned} \langle R_{t2} R_{i1} \rangle &= ((+1)p(R_{t2}) + (-1)p(\overline{R_{t2}})) * ((+1)p(R_{i1}) + (-1)p(\overline{R_{i1}})) \\ &= p(R_{t2})p(R_{i1}) - p(R_{t2})p(\overline{R_{i1}}) - p(\overline{R_{t2}})p(R_{i1}) + p(\overline{R_{t2}})p(\overline{R_{i1}}), \end{aligned}$$

$$\begin{aligned} \langle R_{t1} R_{i2} \rangle &= ((+1)p(R_{t1}) + (-1)p(\overline{R_{t1}})) * ((+1)p(R_{i2}) + (-1)p(\overline{R_{i2}})) \\ &= p(R_{t1})p(R_{i2}) - p(R_{t1})p(\overline{R_{i2}}) - p(\overline{R_{t1}})p(R_{i2}) + p(\overline{R_{t1}})p(\overline{R_{i2}}), \end{aligned}$$

$$\begin{aligned} \langle R_{i1} R_{i2} \rangle &= ((+1)p(R_{i1}) + (-1)p(\overline{R_{i1}})) * ((+1)p(R_{i2}) + (-1)p(\overline{R_{i2}})) \\ &= p(R_{i1})p(R_{i2}) - p(R_{i1})p(\overline{R_{i2}}) - p(\overline{R_{i1}})p(R_{i2}) + p(\overline{R_{i1}})p(\overline{R_{i2}}). \end{aligned}$$

The above products of probabilities are defined as joint probabilities between two independent outcomes. The violation of Equation (12) is a sign of entanglement, and the pair of documents may result in one of the aforementioned Bell states (Equation (10), Equation (11)) as have been described above.

5 Experiment Settings

5.1 Dataset

The proposed model is tested on the ImageCLEF2007 data collection [12], the purpose of which is to investigate the effectiveness of combining image and text for retrieval tasks. Out of 60 test queries we randomly picked up 30 ones, together with the ground truth data. Each query describing user information need consists of three sample images and a text description, whereas each document consists of an image and a text description. For every query, we created a subset of 300 relevant and irrelevant documents, which includes firstly all the relevant documents for the query, and the rest being irrelevant documents. The dataset is used for investigating both the Bell states (Equations (10) and (11)). The number of relevant documents per query ranges from 11 to 98.

5.2 Image and Text Representations-Mono-modal Baselines

The late fusion process is based on mono-modal retrieval scores. For the visual information, feature extraction consists of using the representations learned by the VGG16 model [18], with weights pre-trained on ImageNet to extract features from images, resulting in a feature vector of 2048 floating values for each image. After feature vector extractions, we compute the similarity scores between a submitted visual query and images in the dataset based on Cosine function. For textual information, a query expansion approach has been applied extending the query with the ten most frequent terms according to the ground truth text-based documents. This indeed corresponds to a simulated explicit relevance feedback scenario. Then, the TF-IDF vector representation is used for calculating the text-based Cosine similarity between the a query and text documents. Cosine similarity is particularly used in positive space, where the Cosine similarity score is bounded in $[0,1]$. In our case, we make use of Cosine similarity score for approximating the probability of relevance.

5.3 Experimental Procedure

At the first step, for both text-based and image-based modalities, the Cosine function is employed to approximate the probability of relevance according to a multi-modal query (Fig 5). Then, we create pairs of relevant documents. In the next step, expectation values are computed based on probabilities of relevance according to the process being described in Section 4. The probability for a document to be relevant concerning a modality is equal to the result of Cosine function. Consequently, the probability for a document to be irrelevant concerning the same modality equals 1 minus the result of Cosine function. Then, we fit the CHSH inequality with the calculated expectation values and check for any existence of violation. For each query, we calculate in total the percentage of documents show a violation of the CHSH inequality. At the end of the experiment, we calculate the percentage of queries showing violation.

6 Results and Discussion

The experiment results are out of our expectations since we did not observe any violation of Bell’s inequality. This implies that in the context of our experimental setting non-classical correlations between pairs of documents may not exist, but also that the hypothesis of rotation invariance falls down. Thus, the image-based and text-based bases are not equal Bell states as defined in Equation (4).

This result may be related to our experimental setting that the outcomes of the observables are initially independent. For instance, the probability of the text-based relevance of the first document does not affect the probability of the text-based relevance of the second document. Thus, the joint probability of relevance is calculated as a product of individual relevance probabilities. However, in [1,2,6,7] the Bell inequality has been violated. In those experiments, the users are asked to report their judgments on composite states. Hence the joint probabilities can be directly estimated from the judgments. Thus, the expectation values are calculated under an implicit assumption that the outcomes can be incompatible. This assumption may result in “conjunction fallacy” [20] violating the monotonicity law of probability by overestimating the joint probability, thus violating the Bell inequality.

Our result may be also due to the dataset that has been used to conduct the experiment. In ImageClef2007, the outcomes are independent, i.e., the text-based and image-based relevance, therefore we cannot make the opposite assumption. Thus, we may need another dataset containing relevance judgment for a pair of documents. Additionally, we may search for a dataset where Bell states (i.e., Equation (2)) preexist, such that an interaction between two documents cannot be validly decomposed and modeled as interaction of separate documents. Then, the Bell inequality may be violated for those cases.

Finally, we experimentally investigated the violation of the Bell inequality in a small-scale experiment. In the current experiment, for each query, we focused on a small amount of relevant and irrelevant multimodal documents trying to search for non-classical correlations between two documents. However, it is worth conducting a large-scale experiment as well, looking also at a general first round retrieval process, or even at relevance feedback scenario. Moreover, it would be interesting to investigate the existence of non-classical correlations among many documents. Then, the CHSH inequality should be generalized for systems with multiple settings or basis [10].

7 Conclusion

In this paper, we have investigated non-classical correlations between pairs of decision fused multimodal documents. We examined the existence of such correlations through the violation of the CHSH inequality. In this case, a violation implies that measuring a mono-modal decision in a document, we could instantaneously predict with certainty a mono-modal decision in the other system acquiring information about how to fuse local decisions. Unfortunately, we did

not find any violation of the Bell inequality. This result may be related to our assumption that the outcomes of the observables are initially independent. The result may also be due to the dataset. On one hand there is no real user involved in relevance judgment; on the other hand there do not exist initial Bell states between two multimodal documents. Nevertheless, the experimental results and discussions may provide theoretical and empirical insights and inspirations for future development of this direction.

A Appendix

The expectation of a random variable X that takes the values $\{+, -\}$ according to the probability distribution $P_{X(+)}, P_{X(-)}$ is defined as

$$\langle X \rangle = (+)P_{X(+)} + (-)P_{X(-)}.$$

For two random variables X, Ψ , that take the values $\{+, -\}$ according to the probability distribution $P_{X(+)}, P_{X(-)}$ and $P_{\Psi(+)}, P_{\Psi(-)}$ respectively, the expectation value is defined as the product resulting in

$$\begin{aligned} \langle X, \Psi \rangle &= ((+)P_{X(+)} + (-)P_{X(-)}) * ((+)P_{\Psi(+)} + (-)P_{\Psi(-)}) \\ &= (+)(+)(P_{X(+)}P_{\Psi(+)} + (+)(-)(P_{X(+)}P_{\Psi(-)}) \\ &\quad + (-)(+)(P_{X(-)}P_{\Psi(+)} + (-)(-)(P_{X(-)}P_{\Psi(-)}). \end{aligned}$$

Acknowledgement

This work is funded by the European Unions Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 721321.

References

1. Aerts, D., Sozzo, S.: Quantum structure in cognition: Why and how concepts are entangled. In: International Symposium on Quantum Interaction. pp. 116–127. Springer (2011)
2. Aerts, D., Sozzo, S.: Quantum entanglement in concept combinations. International Journal of Theoretical Physics **53**(10), 3587–3603 (2014)
3. Aspect, A., Grangier, P., Roger, G.: Experimental realization of einstein-podolsky-rosen-bohm gedankenexperiment: a new violation of bell’s inequalities. Physical review letters **49**(2), 91 (1982)
4. Atrey, P.K., Hossain, M.A., El Saddik, A., Kankanhalli, M.S.: Multimodal fusion for multimedia analysis: a survey. Multimedia systems **16**(6), 345–379 (2010)
5. Baltrušaitis, T., Ahuja, C., Morency, L.P.: Multimodal machine learning: A survey and taxonomy. IEEE Transactions on Pattern Analysis and Machine Intelligence (2018)
6. Bruza, P.D., Kitto, K., Ramm, B., Sitbon, L., Song, D., Blomberg, S.: Quantum-like non-separability of concept combinations, emergent associates and abduction. Logic Journal of IGPL **20**(2), 445–457 (2011)

7. Bruza, P.D., Kitto, K., Ramm, B.J., Sitbon, L.: A probabilistic framework for analysing the compositionality of conceptual combinations. *Journal of Mathematical Psychology* **67**, 26–38 (2015)
8. Cirel’son, B.S.: Quantum generalizations of bell’s inequality. *Letters in Mathematical Physics* **4**(2), 93–100 (1980)
9. Clauser, J.F., Horne, M.A., Shimony, A., Holt, R.A.: Proposed experiment to test local hidden-variable theories. *Physical review letters* **23**(15), 880 (1969)
10. Gisin, N.: Bell inequality for arbitrary many settings of the analyzers. *Physics Letters A* **260**(1-2), 1–3 (1999)
11. Gleason, A.M.: Measures on the closed subspaces of a hilbert space. *Journal of mathematics and mechanics* pp. 885–893 (1957)
12. Grubinger, M., Clough, P., Hanbury, A., Müller, H.: Overview of the imageclef-photo 2007 photographic retrieval task. In: *Workshop of the Cross-Language Evaluation Forum for European Languages*. pp. 433–444. Springer (2007)
13. Hou, Y., Song, D.: Characterizing pure high-order entanglements in lexical semantic spaces via information geometry. In: *International Symposium on Quantum Interaction*. pp. 237–250. Springer (2009)
14. Hou, Y., Zhao, X., Song, D., Li, W.: Mining pure high-order word associations via information geometry for information retrieval. *ACM Transactions on Information Systems (TOIS)* **31**(3), 12 (2013)
15. Melucci, M.: *Introduction to information retrieval and quantum mechanics*, pp. 156–158, 176–181, 212–213, 217–221. Springer (2015)
16. Nielsen, M.A., Chuang, I.: *Quantum computation and quantum information* (2002)
17. Pathak, A.: *Elements of quantum computation and quantum communication*, pp. 92–98. Taylor & Francis (2013)
18. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014)
19. Stenger, V.J.: *Timeless reality: symmetry, simplicity and multiple universes*, chap. Ch. 12
20. Tversky, A., Kahneman, D.: Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological review* **90**(4), 293 (1983)
21. Van Rijsbergen, C.J.: *The geometry of information retrieval*. Cambridge University Press (2004)
22. Veloz, T., Zhao, X., Aerts, D.: Measuring conceptual entanglement in collections of documents. In: *International Symposium on Quantum Interaction*. pp. 134–146. Springer (2013)