Boolean Logic Algebra Driven Similarity Measure for Text Based Applications

Hassan I. Abdalla[1], Ali A. Amer[2]

[1] College of Technological Innovation, Zayed University, P.O. Box 144534 Abu Dhabi, U.A.E,
Hassan.Abdalla@zu.ac.ae
[2] Computer Science Department, Taiz University, Yemen, aliaaa2004@yahoo.com

Corresponding Author:
Ali A. Amer[2]
Computer Science Department, Taiz University, Yemen,
Email address: aliaaa2004@yahoo.com; aliaaa2004@gmail.com

ABSTRACT

In Information Retrieval (IR), Data Mining (DM), and Machine Learning (ML), similarity measures have been widely used for text clustering and classification. The similarity measure is the cornerstone upon which the performance of most DM and ML algorithms is completely dependent. Thus, till now, the endeavor in literature for an effective and efficient similarity measure is still immature. Some recently-proposed similarity measures were effective, but have a complex design and suffer some inefficiency. This work, therefore, develops an effective and efficient similarity measure of a simplistic design for text-based applications. The measure developed in this work is driven by Boolean logic algebra basics (BLAB-SM), which aims at effectively reaching the desired accuracy at the fastest run time as compared to the recently developed state-of-the-art measures. Using the term frequency–inverse document frequency (TF-IDF) schema, the K-nearest neighbor (KNN), and the K-means clustering algorithm, a comprehensive evaluation is presented. The evaluation has been experimentally performed for BLAB-SM against seven similarity measures on two most-popular datasets, Reuters-21 and Web-KB. The experimental results illustrate that BLAB-SM is not only more efficient but also significantly more effective than state-of-the-art similarity measures on both classification and clustering tasks.

INTRODUCTION

Over the last ten years, Natural Language Processing (NLP) has been subjected to rapid development as new techniques and methods are continuously introduced to satisfy the ever-disseminating availability of data. This kind of development is clearly being mostly reflected in the fields of Information Retrieval (IR), Data Mining (DM), and Machine Learning (ML) in which several techniques used the similarity/distance measures for different purposes (Zhang & Zuo,

40 2019; Gweon, Schonlau & Steiner, A. 2019; Amer 2020; Teboulle & Nicholas, 2005; Nguyen &
41 Amer, 2019).
42
43 Several machine learning techniques have recorded a surpassing performance, in the NLP field, to
44 handle the voluminous constantly-piling data and information on the internet. Among these
45 techniques are clustering and classification which are still commonly used in almost all scientific
46 fields including text mining, information retrieval, web search, pattern recognition, and biomedical
47 based text mining (Amer & Abdalla, 2020; Rachkovskij, 2017; Gweon, Schonlau & Steiner, B.
48 2019; Kanungo, Mount, Netanyahu, Piatko, Silverman & Wu, 2019; Holzinger, Schantl,
49 Schroettner, Seifert & Verspoor, 2014). For example, in (Holzinger, Schantl, Schroettner, Seifert
50 & Verspoor, 2014), a detailed survey in biomedical-based text mining and classification was done,
51 while stressing the importance of involving and improving similarity measures for classification
52 tasks. Literature has long been stressing the performance of text clustering and classification which
53 depends mainly on the similarity measures. Essentially, both tasks need highly-effective and
54 maximally-efficient similarity measures to reach the desired rendering. However, finding suitable
55 similarity measures for text clustering and classification is a challenging task. Efficiency and
56 effectiveness are the basic characteristics each similarity measure should enjoy.
57
58 Generally speaking, in information retrieval, the documents are drawn as vectors in the vector
59 space model (VSM) (Amer & Abdalla, 2020). In each document's vector, each cell refers to the
60 value of the relative feature that corresponds to the term presence/absence. In this vector, this value
61 is integer/real number represented by the weighting schema which is commonly falling into one
62 of three cases: (1) term frequency (used in BoW) which indicates the appearance times of the
63 relative term, (2) relative term frequency (TF) which is computed as the ratio between the
64 frequency of term and the net number of appearances of all unique terms in the whole set of
65 documents, (3) TF-IDF which is the most commonly-used schema in information retrieval (Afzali
66 & Kumar, 2017). Occasionally, documents are broken into a large number of features represented
67 in VSM, resulting in high sparse VSM. In other words, VSM would contain a rather low percentage
68 of non-zero feature values. Consequently, the sparsity along with the dimensionality curse could
69 have a severely negative impact on the performance of classifiers.
70
71 To tackle such challenges, in IR literature, a dozen of works have introduced several effective
72 similarity measures for text clustering and classification (Amer & Abdalla, 2020; Oghbaie &
73 Mohammadi Zanjireh, 2018; Sohangir ., & Wang, 2017; Lin, Jiang & Lee, 2014; Shahmirzadi,
74 Lugowski, & Younge, 2019; Ke, 2017; White & Jose, 2004; Lakshmi & Baskar, 2021; Kotte,
75 Rajavelu & Rajsingh, 2020; Thompson, Panchev & Oakes, 2015). However, except for (Amer &
76 Abdalla, 2020), these studies proposed similarity measures without providing sufficient insights
77 into run-time efficiency. In other words, these studies might introduce effective measures yet time-
78 inefficient. Moreover, these measures, which are shown effective (Amer & Abdalla, 2020;
79 Oghbaie & Mohammadi Zanjireh, 2018; Sohangir & Wang, 2017; Lin, Jiang & Lee, 2014;
80 Shahmirzadi, Lugowski, & Younge, 2019; Lakshmi & Baskar, 2021; Robertson, 2004), suffer the

81 design complexity. Motivated by this, this work comes with the ultimate aim of finding an
82 influential solution to the efficiency as well as the design complexity of those similarity measures.
83 A Boolean logic algebra-driven similarity measure (BLAB-SM) is simply designed with the aim
84 of its being significantly effective and highly efficient. BLAB-SM takes the presence and absence
85 of each term as long as this term exists in either or both documents under consideration, making it
86 highly competitive to give a robust classification. Seven similarity measures are thoroughly
87 examined under diversified conditions. Using TF-IDF representation (Zhao & Mao, 2018; Joulin,
88 Grave, Bojanowski & Mikolov, 2017), the K-nearest neighbor (KNN), and the K-means algorithm,
89 this work investigates each similarity measure with varying both the K values of KNN and the
90 number of features of each dataset. Similarly to (Amer & Abdalla, 2020) that used the BoW model,
91 we have used the TF-IDF model to assess all measures against low dimensional datasets by
92 analyzing performance on (50, 100, 200, and 350 features) and high dimensional datasets (3000,
93 6000, and all features of each dataset). The performance of measures was studied profoundly to
94 specify the measures that would yield the desired results in each K value of KNN and on each
95 number of features. The key contributions, of this work, are drawn as follows:

97 1. Presenting a new similarity measure whose behavior is driven from the Boolean
98   logic algebra mechanism. It is named Boolean logic algebra-based similarity
99   measure for text clustering and classification (BLAB-SM). Based on a rigorous
100   experimental study, BLAB-SM has been shown a top performer by outperforming
101   all the state-of-the-art measures concerning both effectiveness and efficiency.
102   Indeed, BLAB-SM is one of the fastest measures comparing with all considered
103   measures in this study including cosine, Euclidean, and Manhattan. The
104   experimental results illustrate that BLAB-SM is not only more efficient but also
105   significantly more effective than state-of-the-art similarity measures on both
106   classification and clustering tasks.

107

108 2. Drawing a comprehensive and fair evaluation study for BLAB-SM against seven
109   similarity measures. The performance of all similarity measures is benchmarked on
110   web-KB and Reuters-21 datasets. Using TF-IDF, the KNN classifier, and the K-
111   means clustering, a comparative study for these measures, regarding the
112   effectiveness and efficiency, is made. Interestingly, this study has experimentally
113   shown that the Jaccard similarity measure is an ineffective measure for TF-IDF-
114   based document matching.

115

116 **Paper organization**
117 The related works, including the compared methods in this work, are covered in "RELATED
118 WORK". "THE PROPOSED METHOD" introduces the proposed similarity measure.
119 Experimental settings are concisely presented in "EXPERIMENTAL SETUP". The outcomes of
120 experimental study are drawn in "RESULTS". The most important points out of this study are

121 articulated in "DISCUSSION". Finally, "CONCLUSIONS AND FUTURE WORK" concludes
122 the paper, and presents the avenues of future work.
123

RELATED WORK

125 In IR literature, the vector space model (VSM) has widely been utilized to find the pairwise
126 document similarity using the relative similarity measures. For example, as geometric measures,
127 the Euclidean, Minkowski, Manhattan, and Chebyshev distances are utilized in (VSM) for text
128 classification through finding the distance between each vector pair (Heidarian & Dinneen, 2016;
129 Cordeiro, Amorim & Mirkin, 2012). In general, Manhattan distance is mostly more efficient than
130 Euclidean on small datasets, yet it has long been recorded to have less accurate results in most of
131 the text classification tasks particularly with THE sparse data. In our work, Euclidean was seen
132 more efficient when run on web-kb and Reuters, though, That is due to the Manhattan is being
133 contingent upon the rotation of the coordinate system, leading to its being disadvantageous for
134 both document classification and clustering tasks (Kumar, V. , Chhabra & Kumar, D, 2014).
135 Meanwhile, Jaccard, Ex-Jaccard, Kullback–Leibler divergences (KLD), and Bhattacharya
136 coefficient were all used for the several tasks of ML and IR including text clustering and
137 classification (Amer & Abdalla, 2020; Tanimoto, 1957; Tata & Patel, 2007, Oghbaie &
138 Mohammadi Zanjireh, 2018, François, Wertz, & Verieysen, 2007; D'hondt, Vertommen,
139 Verhaegen, Cattrysse & Duflou, 2010; Li, Yao, Fan & Yu, 2017; Kullback & Leibler, 1951).
140

141 On the other hand, Jaccard is always used to compute the ratio of the number of terms/points
142 included in the document pair (or feature space) to the number of terms included in, at least, either
143 one document/point. However, the Jaccard similarity measure does not utilize the term frequencies
144 in the vector space model, making it an ineffective option for TF-IDF-based matching. Ex-Jaccard,
145 therefore, comes to tackle the limitation of Jaccard when dealing with TF-IDF. However, Ex-
146 Jaccard has long been shown time-inefficient. Both facts have been experimentally corroborated
147 in our study of this paper. In contrast to most similarity measures/ distance metrics, KL divergence
148 is asymmetric. That is, the KLD value from document d1 to document d2 does not equal the KLD
149 value from d2 to d1. This fact contributes negatively to the performance of KLD when text
150 classification considered. All of these distance/similarity measures are considered among the most
151 efficient methods in IR and ML fields; yet, their performance has not reached the desired
152 effectiveness, or even poor chiefly when the data are sparse or of high dimensions or both
153 combined (Subhashini & Kumar, 2010; Li & Han, 2013). On the other hand, Cosine similarity,
154 like Euclidean distance, is one of the most widely-applied similarity measures for text clustering
155 and classification (Amer & Abdalla, 2020; Arivarasan & Karthikeyan, 2019; Zhao & Karypis,
156 2002). It seeks to find the cosine angle between vectors of each document pair. Cosine has long
157 been shown to be effective and efficient at the same time. However, cosine suffers two limitations:
158 (1) its effectiveness and efficiency have been degraded drastically when applied on high-
159 dimensional datasets, or run on datasets of overlapping classes (Amer & Abdalla, 2020), (2) Cosine
160 is more suitable for text mining when data are a symmetric. Otherwise, Cosine might be either
161 biased or less accurate (Afzali & Kumar, 2017).

162

163 To tackle this limitation, therefore, IR and ML literature has still been introducing new similarity
164 measures. Off the most recently-published similarity measures are the set theory-based similarity
165 measure (STB-SM). STB-SM was proposed in (Amer & Abdalla, 2020) for text classification and
166 clustering, and compared against seven similarity measures in the context of the bag of word
167 model. STB-SM was proven maximally effective and significantly efficient. In (Heidarian &
168 Dinneen, 2016), a geometric measure was introduced to find the similarity degree between each
169 document pair. Depending on the information theory (D'hondt, Vertommen, Verhaegen, Cattrysse
170 & Duflou, 2010; Aslam & Frost, 2003), an Information-Theoretic measure for document Similarity
171 (IT-Sim) was proposed. In the same line, in (Sohangir & Wang, 2017), a similarity measure, named
172 Improved Sqrt-Cosine similarity (ISC), was developed for text classification, and proven effective.
173 In the same context, the pairwise document similarity measure (PDSM) was proposed in (Oghbaie
174 & Mohammadi Zanjireh, 2018). PDSM took into account the feature weights and the number of
175 features found in, at least, one document. Another popular similarity measure for text processing,
176 SMTP, was presented in (Lin, Jiang & Lee, 2014) and shown highly effective when run against
177 the state-of-art similarity measures including cosine, Euclidean and Sim-IT.

178

179 Even though the literature has presented a good number of similarity measures, the problem is that
180 these proposed measures were either seen effective like PDSM, SMTP, and ISC; yet, they were
181 time-inefficient. On the other extreme, other measures were seen efficient like Euclidean and
182 Manhattan, yet are not as effective as PDSM, ISC or SMTP. As a compromised solutions, Cosine
183 and STB-SM were described with their being closely effective and reasonably time-efficient. SYB-
184 SM was seen much more effective (Amer & Abdalla, 2020), though. Given the drawn-above
185 limitations of previous similarity measures, our work of this paper endeavors to introduce a novel
186 similarity measure that would tackle these limitations while shrinking the trade-off between
187 efficiency and effectiveness to the greatest extent. The proposed measure behaves based on
188 Boolean logic algebra basis, called BLAB-SM, has been developed, and experimentally shown to
189 outperform all compared state-of-art effectively and efficiently.

190

191 **The Compared Methods**
192 In the following, all compared similarity measures and distance metrics are described. Having two
193 documents document1 and document2 whose TF-IDF of their "n" terms (w.t) have been saved in
194 both vectors: $doc_1(w.t_{11}, w.t_{12, ...}, w.t_{1n})$ and $doc_2(w.t_{21}, w.t_{22, ...}, w.t_{2n})$ respectively, the considered
195 seven similarity measures or distance metrics [29-32] are listed as follows;

196

197 **Euclidean Distance (ED)**

198 ED computes the distance between each point pair in N-dimensional space. It is define by the
199 following equation:

200 $$D_{Euc}(doc1, doc2) = \sum \sqrt{(doc_{11\_}doc_{12})^2 + (doc_{21\_}doc_{22})^2 + ... (doc_{n1\_}doc_{n2})^2} \qquad (1)$$

201

202   **Cosine Measure**

203   The pairwise similarity is found between each document pair using both the dot product as well as
204   the magnitude of both vectors doc_1 and doc_2.

205    $Sim_{Cos}(doc1, doc2) = \dfrac{\sum\limits_{i=1}^{n} doc_{i1} * doc_{i2}}{\sqrt{\sum\limits_{i=1}^{n} doc_{i1}^2} * \sqrt{\sum\limits_{i=1}^{n} doc_{i2}^2}}$ (2)

206
207   **Jaccard Similarity Measure**

208   It is one of the most widely used similarity measures. However, the problem with Jaccard as we
209   found experimentally is that Jaccard is highly reliant on the common values which are, unlike BoW
210   representation in which Jaccard is working well, barely exist in TF-IDF representation. This
211   reliance factor makes Jaccard a very poor option when used for TF-IDF-based document matching.
212   This measure is defined by next equation.

213    $Sim_{jaccard}(doc1, doc2) = \dfrac{doc1 \cap doc2}{doc1 \cup doc2}$ (3)

214
215   **Extended Jaccard**

216   The extended Jaccard (ex-jaccard, in short) can be considered as the inverse of Jaccard coefficient,
217   and its equation is drawn as follows;

218    $Sim_{Ex-jaccard}(doc1, doc2) = \dfrac{doc1 \,.\, doc2}{|doc1|^2 + |doc2|^2 - (doc1 \,.\, doc2)}$ (4)

219
220   **Manhattan**

221   This distance metric computes the sum of absolute differences between the coordinates of vectors
222   of document pair. It is defined as follows:

223    $Manhattan - distance\ (doc1, doc2) = \sum\limits_{i=1}^{n} |doc1_{i1} - doc2_{i2}|$ (5)

224
225   **Kullback Leibler Divergence (KLD)**
226   kld finds the difference between the probability distributions, and is defined as follow;

227    $Sim_{KL}(doc1, doc2) = \sum\limits_{i=1}^{n} Pdoc1(doc_{i1}) * \log\left(\dfrac{Pdoc1(doc_{i1})}{Pdoc2(doc_{i2})}\right)$ (6)

228   where $Pdoc_1(doc_{i1})$ is the value of ith feature of doc_1 and $Pdoc_2(doc_{i2})$ is the value of ith feature of
229   doc_2.

230

231
232   **SMTP**

233 This similarity measure was presented in (Lin, Jiang & Lee, 2014) for text clustering and
234 classification. It considers two key scenarios when the feature exists in both documents doc_1 and
235 doc_2, and when the feature is absent in the pair. It was defined as follows;

236

237 $$Sim(doc1,doc2) = \frac{F(doc1,doc2) + \lambda}{1 + \lambda} \qquad\qquad (7)$$

238 where

239 $$F(doc1,doc2) = \frac{N_{star}(doc1,doc2)}{N_{union}(doc1,doc2)}$$

240

241 $$N_{star}(doc1,doc2) = \begin{cases} 0.5\left(1 + exp^{-(\frac{doc_{i1} - doc_{i2}}{var})^2}\right) & if\ doci1.doci2 > 0 \\ 0\ if\ doc_{i1} = 0, doc_{i2} = 0 \\ -\lambda\ \ Otherwise \end{cases}$$

242 And

243 $$N_{union}(doc1,doc2) = \begin{cases} 0,\ if\ doci1 = 0, doci2 = 0 \\ 1,\ \ Otherwise \end{cases}$$

244

245 Where λ is a constant that was fixed at values of (1) and (0.01), while conducting our experiments,
246 for classification and clustering respectively. The variable var indicates the standard distribution
247 of all non-zero features. The values of (1) and (0.01) were not randomly chosen for λ. They were
248 selected based on the setting asserted and used in (Lin, Jiang & Lee, 2014) as the optimal option
249 to compare SMTP with other measures for both classification and clustering.

250

251 ## THE PROPOSED METHOD

252

253 **Motivation**
254 As mentioned earlier in the drawn-above sections, the text classification and clustering is tricky
255 task chiefly when big data is set to be handled. Furthermore, even though the literature is full of
256 effective similarity measures that have a good performance, the necessity for efficient and effective
257 similarity measures is still incomplete. While there have been some efficient measures like
258 Euclidean and Manhattan, these measures have been shown of poor performance when applied on
259 texts of middle, big, or voluminous-sized datasets. On the other hand, there have been efficient
260 and effective measures like cosine; however, cosine measure does not reach the desired
261 performance. Therefore, several works in literature have been presenting new effective measures
262 like PDSM, ISC, SMTP, and STB-SM. While PDSM was seen as a rather time-inefficient measure
263 (Amer & Abdalla, 2020), SMTP has been seen time-inefficient as well in our work. On the other
264 hand, these measures were seen much more effective than cosine; Yet not as efficient as cosine.
265 The need to address the trade-off between efficiency and effectiveness motivates us mainly to
266 enrich IR and ML with a new efficient and effective similarity measure (BLAB-SM). The newly-
267 proposed measure has been proposed to shrink the already-mentioned trade-off. Experimentally,

268 BLAB-SM is seen as effective as SMTP, and more efficient than cosine. It is also seen as efficient
269 as Euclidean and Manhattan in most cases which is the ultimate objective of this work. Concisely,
270 BLAB-SM comes to effectively fill the gap of efficiency problem from which some effective
271 measures like SMTP suffers, while maintaining, if not outperforming in some cases, the
272 effectiveness reached by SMTP.
273

274 **BLAB-SM Similarity Measure**

275 Three basic definitions for the Boolean logic algebra, which inspired us to propose the measure,
276 are briefly highlighted before presenting our proposed measure, BLAB-SM. In doing so, BLAB-
277 SM's explanation and analysis would be completely understood.
278

279 The Logical Gates

280 Strictly speaking, the digital systems are constructively defined by using the logic gates. In general,
281 the basic gates are the AND, OR, NOT gates. The basic operations are described below along with
282 their truth tables as follows;
283

284   Figure 1 AND Gate   Table 1 A AND B Truth Table
285
286

287 The AND gate is an electronic circuit that yields an output (1) only and only if all its inputs (A and
288 B, in Figure 1, Table 1) are (1). This symbol ($\wedge$) is used to show the AND operation, i.e. A $\wedge$ B.
289 AND gate in our measure comprises only and only the shared features.
290

291   Figure 2 OR Gate   Table 2 A OR B Truth Table
292

293 Like AND gate, the OR gate is also an electronic circuit. However, the OR gate yields an output
294 (1) if and only if one or more of its inputs (A and B, in Figure 2, Table 2) are (1). This symbol ($\vee$)
295 is used to show the OR operation, i.e. A $\vee$ B. OR gate in our measure concerns both the shared
296 and non-shared features at the same time.
297

298   Figure 3 EXOR Gate   Table 3 A EXOR B Truth Table
299

300 The 'Exclusive-OR' gate is a circuit that yields an output (1) if either one, but not both, of its two
301 inputs are (1). The symbol with an encircled plus sign ($\oplus$) is utilized to signal the Ex-OR operation,
302 i.e. A $\oplus$ B (see A and B in Figure 3 and Table 3). EXOR gate in our measure concerns only and
303 only the non-shared features so as to non-shared features are given a chance to contribute in
304 calculating similarity degree.
305

306 **The Boolean Logic Algebra Based Similarity Measure (BLAB-SM)**

307 Based on the already drawn concepts, the Boolean algebra based similarity measure (BLAB-SM)
308 is defined as follows;
309      $BLAB-SM\ (doc1, doc2) = \alpha \times X + \beta \times Y$ (8)

310

311   Both α and β are parameters that have experimentally been tuned, as drawn in the results section,
312   just to find their best values to detect the highest possible similarity. Although X is concerned with
313   finding similarities between all non-shared features, it finds similarity between all shared features
314   implicitly as well. It finds similarities based on logical sensing of the differences between both
315   documents under consideration. All documents are logically treated and processed the same way
316   logical gates work. On the other hand, Y emphasizes similarity through finding likeness between
317   documents across all shared features only. Assuming having two documents which are document1
318   and document2 which are represented by vectors doc_1 and doc_2 in the vector space model using
319   TF-IDF representation. Using doc_1 and doc_2 with "n" terms, X and Y are defined as follows;

320
$$X = \left(1 - \frac{\sum_{I=1}^{n}(doc_{i1} \oplus doc_{i2})}{\sum_{I=1}^{n}(doc_{i1} \vee doc_{i2})}\right)$$

321
$$Y = \left(\frac{2 \times (\sum_{I=1}^{n}(doc_{i1} \wedge doc_{i2}))}{\left(\sum_{I=1}^{n}doc_{i1}\right) + \left(\sum_{I=1}^{n}doc_{i2}\right)}\right)$$

322

323   Meanwhile, doc_1 or doc_2=1, if doc_1 or $doc_2 \geq 1$; Otherwise, 0.

324

325   **BLAB-SM Analysis**

326   In this sub-section, we concisely, simply and informatively analyze the cases of the proposed
327   measure. Assuming that X consists of X_1 and X_2, and Y consists of Y_1, Y_2, and Y_3. Accordingly, the
328   X and Y equations are re-drawn as follows;

329
$$X = \left(1 - \frac{X_1}{X_2}\right)$$

330   where:

331
$$X_1 = \sum_{I=1}^{n}(doc_{i1} \oplus doc_{i2})$$

332

333
$$X_2 = \sum_{I=1}^{n}(doc_{i1} \vee doc_{i2})$$

334   Similarly:

335
$$Y = \left(\frac{2 \times Y1}{Y2 + Y3}\right)$$

336 where:

$$Y_1 = \sum_{I=1}^{n}(doc_{i1} \wedge doc_{i2}),$$

338

$$Y_2 = \sum_{I=1}^{n} doc_{i1}, \text{ and } Y_3 = \sum_{I=1}^{n} doc_{i2}$$

339
340 The perfect dissimilarity case:
341 This case happens when:
342 $\qquad X_1 = X_2 \Rightarrow X = \text{zero}$, because X would equal $(1 - 1 = 0)$
343 $\qquad \because X_1 = X_2 \Rightarrow Y_1 = zero$ (it is here AND gate), $X_1 = 1$ since $\frac{X_1}{X_2} = 1$
344 $\qquad \Rightarrow Y = \frac{2\, x\, zero}{Y_2 + Y_3} = zero$, regardless of Y2 and Y3 values
345 $\qquad \Rightarrow \text{BLAB} - \text{SM}(doc1, doc2) = \alpha \times X + \beta \times Y = \alpha \times zero + \beta \times zero = zero$
346

347 Example (perfect dissimilarity case): assuming we have doc1 (3, 0, 1) and doc2 (0, 2, 0). By the
348 applying the perfect dissimilarity scenario, we find that BLAB-SM=zero, for both documents (1,
349 0, 1) and (0, 1, 0), which is logically true since there is no shared feature exist.
350
351 The average similarity:
352 $\qquad X_2 > X_1 > zero \Rightarrow Y_1 > zero \Rightarrow 1 > X > zero$, and $1 > Y > zero$
353 $\qquad\qquad\qquad\qquad \Rightarrow 1 > \text{BLAB} - \text{SM}(doc1, doc2) > zero$
354 Where (1) is the upper bound and (0) is the lower bound.
355

356 Example (Average similarity): assuming we have doc1 (3, 1, 1) and doc2 (6, 2, 0). By applying
357 the average case scenario on both documents (1, 1, 1) and (1, 1, 0), BLAB-SM would have a
358 value of roughly (0.73) which is bigger than zero and less than 1. It is worth indicating that 0.73
359 is logically reasonable than the cosine value which would reach 0.95. In fact, this is one novelty
360 of our measure as similarity has never been exaggerated like what is done with the most state-of-
361 art measure. Our measure allows non-zero non-shared features to have an implicit contribution to
362 the similarity computation. Therefore, BLAB-SM takes the presence and absence of all features
363 into consideration effectively.
364
365 The perfect similarity case:

$$X_1 = zero \Rightarrow X = 1 - \frac{zero}{X_2},$$

367 $\qquad X = 1$, regardless of the value of X_2
368 $\qquad \because X_1 = zero \Rightarrow Y_1 = \overline{X_1},$
369 $\qquad \Rightarrow Y_1 = 1$, since $Y_1 = complement\,(X_1) \Rightarrow Y = \frac{2\, x\, 1}{Y_2 + Y_3}$
370 $\qquad \because Y_1 = 1 \Rightarrow Length\,(doc1) = Length\,(doc2) \Rightarrow Y_2 = Y_3$
371 $\qquad \Rightarrow Y = \frac{2}{2x\, Y_2\,(or\, 2x\, Y_3)} = \frac{2}{2} = 1$
372

373 Example (perfect similarity Case): assuming we have doc1 (3, 3, 3) and doc2 (3, 3, 3), or doc1
374 (1, 1, 1) and doc2 (1, 1, 1). By applying the best case scenario, we find that BLAB-SM=1 which
375 is logically true as both documents are equal and equivalent.
376

377 **Similarity Measure Characteristics**
378   Six characteristics should be defined on each similarity measure so this measure could be
379 classified as good measure (Haroutunian 2011; Amigó, Giner, Gonzalo, & Verdejo, 2020). These
380 characteristics are given as follows;
381

382 **Characteristic 1:** The presence or non-presence of the targeted feature is more important than
383 the difference between the values connected with the present feature.
384

385 Assuming we have doc_1 (3, 1, 1) and doc_2 (6, 2, 0). Then, the binary vectors, (1, 1, 1) and (1, 1,
386 0), of both documents are more important than the integer values linked to the values of features
387 themselves in both documents. Let us take feature3 (f_3), we can say that feature3 has no link with
388 doc_2 while it has a link with doc_1. In such case, both documents are dissimilar with respect to
389 feature 3. So, we can conclude that feature3 has more weight in deciding similarity between doc_1
390 and doc_2 than f_1 and f_2 which are exist in both documents. In fact, BLAM-SM focus in this issue
391 when similarity is being calculated as behavior of BLAM-SM goes in the same way the logic of
392 gates mechanism works.
393

394 **Characteristic 2:** The similarity degree should increase when the difference between the non-
395 zero features values decrease. For example, having feature1 and feature2 as two features (f_1 and
396 f_2) of doc_1 and doc_2, similarity (doc_1, doc_2) while f_1=10 and f_2=5 is higher than the similarity
397 between f_1=18 and f_2=4.
398

399 **Characteristic 3:** The similarity degree should be increased when the number of present features
400 grows. For example, having three vectors of three documents; doc_1 (1, 1, 0), doc_2 (1,1,1), doc_3
401 (1, 0, 0), the similarity (doc_1, doc_2) is far higher than similarity (doc_1, doc_3) due to the difference
402 in the number of present or non-present features between all documents.
403

404 **Characteristic 4:** The document Pairs are highly different to each other if there have not had
405 almost equivalent zero-valued features versus non-zero-valued features. For example, having two
406 vectors of two documents doc_1 (f_1, f_2)=(1,0) and doc_2 (f_3, f_4)=(1,1), $doc_1.f_2$ and $doc_2.f_4$ are the
407 main reason for maximizing the difference between both documents as $f_2 * f_4$=0, and, f_2+f_4>0.
408

409 **Characteristic 5:** the symmetric properties should be met by each similarity measure. For
410 example, the similarity degree between doc_1 (0, 1, 1, 0) and doc_2 (1,1,1,1) must be the same
411 when doc_2(1,1,1,1) and doc_1 (0, 1, 1, 0) are addressed.
412

413   **Characteristic 6:** The value of distribution should contribute to the similarity degree between
414   each document pair.
415

416   ## EXPERIMENTAL SETUP

417   In the following, the experimental settings, algorithms, machine and data sets' descriptions,
418   performance evaluation for both classification and clustering tasks are all elaborated.
419

420   For text clustering and classification applications, the feature extraction/selection and the pre-
421   processing are crucial stages. First of all, text pre-processing was done to make the text process-
422   able. Initially, the text was switched from upper case to lower case, and numbers, punctuations,
423   stop words (common words) were removed. The extra spaces and symbols (like $, %) were also
424   eliminated. For word stemming and text representation, python 3 was used to run the pre-
425   processing using Nltk (Natural language toolkit) library. The ntlk word tokenizer was used for
426   tokenization, and the Lemmatizing was done using the ntlk stem WordNetLemmatizer. Finally,
427   stopword removal was done using the ntlk stopwords. Then, the vector space model (VSM) was
428   utilized to represent features with TF-IDF representation (Robertson 2004). The TF-IDF is the
429   multiplication of term frequency (TF) by the inverse document frequency (IDF), and is defined
430   as follows;

431
$$TF-IDF_{(t,d)} = tf(t,doc) * log\left(\frac{N}{df+1}\right) \qquad (9)$$

432
$$where, tf(t,d) = \frac{count\ of\ term\ t\ in\ document\ doc}{number\ of\ words\ in\ document\ doc}$$

433   Where df is the occurrence of t in all considered documents, and N is the total number of
434   documents.
435

436   **KNN Classifier**
437   In general, KNN examines each test point based on its "K" closest "neighboring" points which
438   are sorted and assumed to be the most similar points. Then, the test point is classified based on
439   the voting technique. It takes the class label of the majority vote among all K neighboring
440   training points in the feature space. The neighborhood is basically determined based on the used
441   distance/similarity measure. Let us assume that d_i is the training set, and d_i^* is the testing set, c is
442   the true class of training set, and c^* is the predicted class for the testing set (c, c^*=…., m) where
443   m is the number of classes. In the training classification process, only the true class c of each
444   training set is used, and during testing, class c^* is predicted for each testing set. On the other
445   hand, when using the one-nearest neighbor rule, the predicted class of the testing set d_i^* is set
446   equal to the true class c of its nearest neighbor, where w_i is the nearest neighbor to d^* at the
447   distance:
448

449
$$d\left(w_i, d_i^*\right) = \min_{j}\{d(w_j, d_i^*)\} \qquad 10$$

450

451 For KNN, the predicted test sample class d_i^* is assigned to the most frequent true class among
452 the K nearest training sets.
453
454 The experimental design flow chart, provided in Figure 4, shows briefly how the KNN classifier
455 was used to perform classification task successfully.
456
457   Figure 4. The experimental Design Flow Chart
458
459
460 **Machine Description**
461 This work has been written in python language, and run on Processor Intel Core i5-3320M (2.6
462 GHz), RAM 4GB with OS Windows 7 (64 bit).
463
464
465
466
467 **Datasets Description**
468 Reuters-R8 (Table 4): it consists of 7674 documents with eight classes, and has 18308 features
469 after it has been pre-processed.
470 Web-kb (Table 5): it consists of 4199 documents with four classes, and has 33,025 features after
471 it has been processed. It is composed of web pages of computer science from universities:
472 Cornell, Texas, Washington, and Wisconsin.
473
474 Both Tables (4) and (5) holds the description of both datasets used in this work. The data in each
475 datasets was split into training and testing in ratio 2:1 (67%: 33%). It is worth indicating that
476 these datasets are publicly available[1].
477
478   Table 4 Reuters-R8 dataset – 8 Classes   Table 5 Web-KB dataset – 4 Classes
479
480 **Classification Evaluation Criterion**
481 **Accuracy**
482 It checks the sample total that are unmistakably classified out of the whole collection.
483   $ACC = \dfrac{\text{True Positive} + \text{True Negative}}{\text{True Positive} + \text{True Negative} + \text{False Positive} + \text{False Negative}}$ (11)
484
485 **Precision**
486 It gives the whole number of items unmistakably identified as positive out of the whole items
487 defined as positive.
488   $PRE = \dfrac{\text{True Positive}}{\text{True Positive} + \text{False Positive}}$ (12)

---

[1] https://gith_ub.com/aliamer/Boolean-Logic-Algebra-Driven-Similarity-Measure-for-Text-Based-Applications/blob/main/Reuters%20%2B%20WebKB%20datasets.rar

489

490 **Recall**

491 It gives the whole number of items unmistakably identified as positive out of the actual positive

492 $REC = \dfrac{\text{True Positive}}{\text{True Positive} + \text{False Negative}}$ (13)

493

494 **F-Method**

495 It is a harmonic mean of precision and recall.

496 $F \text{ Score} = 2 * \dfrac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recal}}$ (14)

497 **G-Method**

498 It is used as a geometric mean of both precision and recall.

499 $GM = \dfrac{\text{True Positive}}{\sqrt{(\text{True Positive} + \text{False Positive}) * (\text{True Positive} + \text{False Negative})}}$ (15)

500

501

502 **Average Mean Precision (AMP)**

503 It is the average of the averaged precision of all classes.

504 $AMP = \sum_{n}(R_n - R_{n-1})P_n$ (16)

505 Where P_n and R_n are the precision and recall at the nth threshold.

506

507

508 **Clustering Evaluation Criterion**

509 **Purity**

510 To check the purity of each cluster as it judges the coherence of a cluster.

511 $$Purity = \frac{1}{N}\sum_{i=1}^{k} max_j |c_i \cap t_j|$$ (17)

512 Where N = number of data points, k = number of clusters, c_i is a cluster in C, and t_j is the

513 classification which has the max count for cluster c_i

514

515 **Entropy**

516 It is used to measure the extent to which a cluster contain single class and not multiple classes.

517 $$EN = \sum_{i=1}^{c} ci * log (ci)$$ (18)

518

519 # RESULTS

520 Before endorsing the best values of parameters α and β, each parameter was thoroughly and

521 carefully tested while its value varied from (0.1) to (0.9) with an increment of (0.1) each time.

522 Varying values from 0.1 to 0.9 for both parameters was to test BLAB-SM on all different

523 scenarios so the desired effect would not be missed. Based on the results of the thoroughly-

524 performed experiments on several versions of BLAB-SM, we approved values (0.5; 0.5) for α
525 and β respectively when running BLAB-SM for both classification and clustering tasks.
526

527 **Classification Results**
528 Fixing α and β on values (0.5; 0.5), this work examines all similarity measures thoroughly using
529 KNN classifier and K-means clustering algorithm. The K value in KNN was changed in each run
530 from (1) to (120) as given in the appendix. Furthermore, the number of features was falling in
531 one of these values (10, 50, 100, 200, 350, 3000, 6000, NF) where NF is the whole number of
532 features. After that, the results are averaged for each measure on all K values as given in Tables
533 (6-14). On the other hand, albeit it has long been proven effective when dealing with the BoW
534 model, the Jaccard measure has been excluded due to its being proven ineffective option when
535 dealing with TF-IDF representation (see the appendix). That is because of the fact that Jaccard is
536 heavily based on the common values between each considered pair. However, the common
537 values with TF-IDF is far less than it is with BoW. This fact justifies the poor behavior of
538 Jaccard when dealing with TF-IDF-based document matching. The bolded values in the next
539 Tables signify the best values. For readability, Accuracy, Precision, Recall, F-Measure, G-
540 Measure, and Average Mean Precision are represented by ACC, PRE, REC, FM, GM, and AMP
541 respectively.
542
543 Table 6. Performance Evaluation, NF=10 – Averaged Results
544

545 In Table 6, for Reuters dataset, the BLAB-SM similarity measure, Manhattan followed by KLD
546 achieved the highest accuracy. Ex-jaccard, although it was not among the best performers but it
547 still outperformed Euclidean, SMTP, and cosine respectively with regard to accuracy. On the
548 other hand, KLD followed by BLAB-SM and Manhattan had the best FM and AMP with BLAB
549 is better with FM. However, with regard to the Web-KB dataset, Manhattan, followed by KLD
550 and Euclidean, met the highest accuracy. For FM and AMP criterions, Manhattan followed by
551 Euclidean and cosine, outweighed all other measures. Thus, the best measures were BLAB-SM,
552 Manhattan, and KLD on Reuters, and Manhattan followed by KLD and Euclidean on Web-KB.
553
554 Table 7. Performance Evaluation, NF=50– Averaged Results
555
556 Table 8. Performance Evaluation, NF=100– Averaged Results
557

558 In Tables 7-8, it is obvious that, for Reuters dataset, BLAB-SM followed by SMTP and cosine,
559 obtained the highest ACC. However, with FM and AMP, Manhattan, Euclidean, cosine and
560 SMTP obtained the best values. On Web-KB dataset, SMTP, followed by BLAB-SM, and
561 cosine, achieved the highest ACC, FM and AMP respectively. Thus, the top performer measures
562 were SMTP, BLAB-SM, and cosine. From Table 9, on Reuters and Web-KB datasets, BLAB-
563 SM, followed by SMTP and cosine, obtained the highest ACC, FM, and AMP with cosine being
564 superior over BLAB-SM and SMTP in terms of FM and AMP. Thus, the top performers were
565 BLAB-SM, SMTP, and cosine.

566

567 Table 9. Performance Evaluation, NF=200– Averaged Results

568

569 In Table 10, on Reuters dataset, Cosine, followed by BLAB-SM and SMTP, obtained the highest
570 ACC. Interestingly, Cosine had been superior to BLAB-SM and SMTP with FM and AMP.
571 However, on Web-KB dataset, like Table 9, BLAB-SM, SMTP followed by cosine obtained the
572 highest ACC, FM, and AMP respectively. Thus, the top performers were BLAB-SM, SMTP, and
573 cosine.

574 Table 10 Performance Evaluation, NF=350– Averaged Results

575

576 From Table 11, on Reuters dataset, SMTP, followed by BLAB-SM and cosine, obtained the
577 highest ACC, FM, and AMP with cosine being superior over SMTP and BLAB-SM in terms of
578 FM and AMP. However, on Web-KB dataset, BLAB-SM, followed by SMTP and cosine,
579 obtained the highest ACC, FM and AMP respectively. Thus, the top performers were BLAB-
580 SM, SMTP and cosine.

581 Table 11 Performance Evaluation, NF=3000– Averaged Results

582

583 In Tables 12-13, on Reuters dataset, SMTP, followed by BLAB-SM and Ex-jaccard, obtained the
584 highest ACC. Cosine, Euclidean, SMTP, and BLAB-SM were the best on FM and AMP.
585 However, on Web-KB dataset, BLAB-SM, followed by SMTP and cosine, obtained the highest
586 ACC, FM, and AMP with cosine being superior over BLAB-SM and SMTP in terms of FM
587 only. Thus, the top performers were BLAB-SM, SMTP, Ex-Jaccard and cosine.

588
589 Table 12. Performance Evaluation, NF=6000– Averaged Results

590
591 Table 13. Performance Evaluation, NF= the whole Number of Features

592

593 Finally, and most importantly in the averaged results case, in Table 14, BLAB-SM and SMTP
594 followed by Cosine were the best measures on both datasets with BLAB-SM taking the lead over
595 SMTP, and SMTP taking the lead over cosine. Surprisingly, cosine/Reuters was the top in terms
596 of FM and AMP.

597 Table 14. Performance Evaluation - the Averaged of Averaged Results

598

599 **Clustering Results**

600 We fixed K on 5, 10, and the number of actual classes of each dataset. The K-means was
601 conditioned to stop after (50) iterations, or the stability has been reached after two successive
602 cycles. Centroids were randomly picked in each iteration. The best measures based on Table 15,
603 BLAB-SM went side by side with Euclidean followed by SMTP and KLD. The bolded values in
604 Tables (15-16) signify the best values.

605 Table 15. Purity

606

607 On the other hand, based on results of Table 16, Manhattan, BLAB-SM, SMTP followed By
608 Cosine have been the best measures.

609 Table 16. Entropy

610

DISCUSSION

612 In this section, we have investigated two key points: (1) the stability of each similarity measure
613 over each dataset, (2) which features number each similarity measure has been the best.

614

615 Measures Stability
616 Recalling the results of Tables 6-14, Table 17 holds the most stable measures. In Table 17, R and
617 W refer to both Reuters and Web-KB respectively.

618 Table 17. Measure Stability Status

619

620 Table 17 shows that SMTP, BLAB-SM and cosine as the most stable measures with 40, 39, and
621 39 points respectively. It gives one plus (++1) for each measure when a measure has been
622 superior in terms of any criterion (ACC, FM, and AMP). On Reuters, the competition was held
623 between BLAB-SM, SMTP, cosine, and Ex-Jaccard. Moreover, it can be concluded, from Table
624 17, that BLAB-SM, SMTP, cosine can be used effectively for both low, middle and high
625 dimensional datasets on each NF. Euclidean and Manhattan also performed well on low
626 dimensional datasets (NF in [10-200] features). EX-Jaccard was observed to behave well on the
627 middle and high dimensional datasets (NF in [350-N] features).

628

Performance Analysis

630 In this section, using the accuracy, f-measure, and average mean precision, we analyze the
631 measure's performance.

632

Reuters

634 Figure 5 shows BLAB-SM, SMTP, Ex-Jaccard, and Cosine achieved the most stable
635 performance. The Cosine and ex-Jaccard showed a punctuated accuracy as cosine was superior
636 when NF was in the range [10-3000]. Ex-Jaccard had the lead as NF grew, though. When
637 NF=6000, competition restricted among Cosine, Ex-jaccard, SMTP, and BLAB-SM with SMTP,
638 BLAB-SM, and Ex-Jaccard being the best. However, as NF exceeded 6000 features, the
639 competition was held between SMTP, BLAB-SM, and Cosine with SMTP and BLAB-SM being
640 the top performers. On the other hand, Manhattan, Euclidean, and kullback Leibler had the worst
641 performance chiefly as NF grew. Albeit that their performance is good when the number of
642 features (NF) were between 10 and 350, performance deteriorated steadily as NF surpasses 350.

643

644 Figure 5 Accuracy over all Measures on all NF values– Average results (K=1-120; +2) – Reuters

645

646 In general, based on Figures 6 and 7, it can be concluded that Cosine, BLAB-SM, and SMTP had
647 an almost close performance with slight superiority given to Cosine. Euclidean was shown a
648 competitor with cosine when all features were considered, though. As expected, KLD,
649 Manhattan, Euclidean had poor performance compared to BLAB-SM, SMTP, and Cosine. The
650 Ex-jaccard had a competitive performance on Reuters, though.

651
652   Figure 6 F-measure over all Measures on all NF values– Average results (K=1-120; +2) – Reuters
653
654
655
656   Figure 7 AMP over all Measures on all NF values– Average results (K=1-120; +2) – Reuters
657

658   **Web-Kb**
659   Figures (8- 10) illustrate the map of the criterion movements (results were averaged) for all
660   measures over several NF values. On one hand, in Figure 8, Cosine had been the middle ground
661   between the first group of measures (KLD, Manhattan, Euclidean, and Ex-Jaccard) and those
662   with the best performance (BLAB-SM and SMTP). BLAB-SM and SMTP had been involved in
663   fierce competition with SMTP being superior as NF was in the range [10-100]. However, BLAB-
664   SM had shown higher superiority as NF grew and surpassed 200 features. On the other hand,
665   Figure 8 reveals that Manhattan and kullback Leibler performed poorly as NF grew.
666   Surprisingly, Ex-Jaccard had poor performance as NF was in the range [10-200]; yet, as NF
667   grew, its performance grew and was competitive with Euclidean and cosine.
668
669   Figures 9 and 10, on the other hand, show almost the same conclusion deduced from figure 7,
670   with BLAB-SM being better than SMTP in terms of FM and AMP. In general, like Reuters, both
671   BLAB-SM and SMTP had a close performance trend with significant superiority given to
672   BLAB-SM regarding ACC and AMP. Interestingly, cosine was highly competitive with BLAB-
673   SM and SMTP in terms of FM and AMP. Euclidean was an equivalent to cosine when all
674   features were considered, though. Finally, KLD, Manhattan, Euclidean, and ex-Jaccard had
675   lower performance on Web-KB compared with BLAB-SM, SMTP, and Cosine. Similarly to
676   Reuters, Euclidean was superior to Ex-Jaccard when NF was in the range [10—350]. However,
677   Ex-Jaccard had the lead when NF grew except for one case.
678
679   Figure 8. Accuracy over all Measures on all NF values– Average results (K=1-120; +2) – Web-KB
680
681   Figure 9 F-measure over all Measures on all NF values– Average results (K=1-120; +2) – Web-KB
682
683   Figure 10 AMP over all Measures on all NF values– Average results (K=1-120; +2) – Web-KB
684
685   Last but not least, Figures 11-12 showcase the averaged performance of all measures on both
686   datasets. Both figures stress that BLAB and SMTP have an almost equal performance trend as
687   the top performers.
688
689   Figure 11. Performance of all Measures on both datasets – Reuters and Web-KB
690
691   Figure 12. Performance of all Measures on both datasets – Reuters and Web-KB
692

693   Finally, to provide the statistical evidence for the robustness of BLAB-SM's performance, the
694   results of BLAB-SM against its rival measures have been verified on both datasets using the
695   standard test of the non-parametric Wilcoxon test and the paired t-test. The statistical details for
696   BLAB-SM, against all seven similarity measures, have been made based on the Accuracy
697   "ACC" metric on both datasets, and the results are drawn in Tables 18-23. Results show that
698   BLAB-SM is highly effective on Web-KB, and competitively effective with SMTP and Ex-
699   Jaccard on Reuters. The statistical analysis is made by setting the standard value of the
700   significance level at 0.05 (95%), and results are analyzed regarding the MAE, MSE, Std Error, z-
701   score, p-value, and t-score. The degrees of freedom (DF) for the paired t-test have been assigned
702   the value of 59.0 as we have 60 K values (1-120, +2) for the KNN classifier (see Appendix). The
703   negative sign of z-score and t-score implies the effectiveness and superiority of BLAB-SM
704   comparing with its rival measures. In other words, both z-score and t-score indicate the
705   significant difference between the BLAB-SM and its rivals.

706   Table 18. Reuters - Statistical Significance of the experimental results of BLAB-SM against all measures
707
708   Table 19. Reuters -Statistical significance with non-parametric Wilcoxon for BLAB-SM against all measures
709
710   Table 20. Reuters -Statistical significance with paired t-test (DF is degrees of freedom) for BLAB-SM against all measures
711
712   Table 21. Web-KB - Significance Test of BLAB-SM against all measures
713
714   Table 22. Web-KB - Statistical significance with non-parametric Wilcoxon for BLAB-SM against all measures
715
716   Table 23. Web-KB - Statistical significance with paired t-test (DF stands for degrees of freedom)
717

718   Furthermore, for both datasets, the p-value of BLAB-SM is less than the value of the
719   significance level (0.05). This is strong evidence for the performance robustness of BLAB-SM
720   comparing with its rival measures. On one hand, based on the statistical results given in Tables
721   18-23, the negative sign of both z-score and t-score shows that the BLAB-SM measure is the top
722   performer comparing with its rival measures. Moreover, z-score values of the BALB-SM
723   measure are maximally significant in all cases as its p-values are less than (0.05) which is the
724   level of significance. On the other hand, on Reuters only, in terms of t-score, all the results are
725   significant except of SMTP and Ex-jaccard. Nevertheless, this slight insignificance of BLAB-
726   SM could be compensated by the maximally-significant run time of BLAB-SM against both
727   SMTP and Ex-Jaccard on Reuters.
728
729   **Clustering Analysis**
730   Using the results of Tables 15-16, the analysis is made in Table 24. The points achieved by each
731   similarity measure are accumulated on each corresponding metric. The whole number of points
732   is 12 points. That is because of having two datasets and two metrics on three values of clustering
733   variable (K=5, K=10, and K equal the number of actual classes). The bolded values in Table 24
734   suggest the highest rank.
735
736   Table 24. Clustering Process - Points and Rank of Similarity Measures

737
738 **Execution Time Analysis**
739 The time of each measure was accumulated and averaged as given in Figures 13-14. It is clear
740 that all measures share one fact, in particular Ex-Jaccard and SMTP, the run time is increasing
741 sharply as NF increases steadily. Figures 13-14 give the run time of classification on Reuters and
742 web-kb respectively.
743
744 Figure 13 Execution Time – Reuters
745
746 As it is shown in Figure 13, BLAB-SM followed by Euclidean and Jaccard were the fastest
747 similarity measures with exception when NF=3000 in which KLD was faster. Manhattan
748 followed by cosine and KLD came the next batch of the fastest measures. On the other hand, the
749 slowest measures were SMTP and Ex-Jaccard with Ex-Jaccard being the slowest. Considering
750 Figure 13, BLAB-SM has been the sole measure that meets the ultimate quest of this work in
751 terms of both effectiveness and efficiency at the same time. It is also worth indicating that the
752 BLAB-SM has been as effective as SMTP and much better than Cosine. However, it is
753 maximally efficient comparing with SMTP and more efficient than Cosine. Considering being
754 faster, BLAB-SM could replace SMTP for text classification and clustering effectively and
755 efficiently.
756
757 Figure 14 Execution Time – Web-KB
758
759 Also, Figure 14 shows that Jaccard, Euclidean, BLAB-SM, and Manhattan were the fastest
760 similarity measures. Interestingly, like Reuters, BLAB-SM confirms it is being fast measure on
761 Web-KB as well. In general, Cosine was faster than KLD yet slower than BLAB-SM. On the
762 other extreme, Ex-Jaccard and SMTP were seen to be slower measures with Ex-Jaccard being
763 the slowest. In conclusion, according to all results drawn in Tables (8-16), and considering
764 Figure 14, BLAB-SM has been the sole measure that meets the ultimate quest of this work in
765 terms of effectiveness and efficiency comparing with SMTP in particular and all measures in
766 general. BLAB-SM has been as effective as SMTP, but maximally efficient compared to SMTP,
767 and it also could replace SMTP and Cosine for text classification effectively and efficiently.
768
769 **Merits of BLAB-SM**
770 In this subsection, it is worth referring briefly to the merits of our proposed measures over its
771 competitors chiefly the SMTP. Concisely, the merits could be drawn as follows; (1) BLAB-SM
772 has a simplistic design compared to SMTP. Moreover, BLAB-SM is bounded by upper and
773 lower values in the same design with no complexity being added in the design, SMTP however
774 needed an additional condition to restrict values of SMTP between zero and one (Kumar
775 Nagwani 2014). (2) BLAB-SM has been shown as effective as SMTP (and even better as it is the
776 case on Web-KB) on the high dimensional dataset as all results asserted this claim when NF
777 grew and exceeded 6000 features on both datasets compared to SMTP. In other words, they both

778 have an almost equal performance trend. Finally, (3) BLAB-SM has been observed to be
779 impressively efficient compared to SMTP in particular and all state-of-art measures in general.
780 Time Tables and graphs asserted this claim (see Table 25). When running measures on all
781 features, the BLAB-SM has barely reached 29 minutes on Reuters and 23 minutes on web-KB
782 compared to SMTP which needed roughly 804.933 minutes on Reuters and 336.567 minutes on
783 Web-KB respectively. It is also noted that BLAB-SM is significantly faster than cosine which
784 needed roughly 58.28 minutes on Reuters and 26.37 minutes on Web-KB respectively. With the
785 BLAB-SM is being high-speed measure of a competitive performance, BLAB-SM could be
786 nominated to replace Cosine and SMTP for the text-based applications, chiefly for the big
787 "large-scale" datasets. The run time has been recorded when all features of both datasets were
788 being considered. Table 25 summarizes the run time of all measures along with the BLAB-SM's
789 improvement Rate against these measures.
790
791 Table 25. Comparison of Classification Run Time (in Minutes) of BLAB-SM against all Measures on both Datasets
792
793
794

795 ## CONCLUSIONS AND FUTURE WORK
796 This work has developed a new similarity measure called Boolean Logic Algebra-Based
797 Similarity measure (BLAB-SM) for the text-based applications. BLAB-SM has been designed to
798 logically treat and process documents the same way the Boolean gates work. In the process of
799 developing our BLAB-SM measure, a thorough experimental study has been conducted for
800 several similarity measures for text classification and clustering. The results conclusively
801 showed that the BLAB-SM similarity measure achieved a highly competitive performance on
802 text retrieval quality (classification and clustering) and run-time efficiency. The experimental
803 results,   revealed that BLAB-SM, SMTP, and Cosine scored the highest performance trends
804 (with BLAB-SM and SMTP being the top performers) compared to Ex-Jaccard, Euclidean,
805 Manhattan, and kullback Leibler measures. Ex-Jaccard showed a competitive performance on
806 Reuters, though. On the other hand, Manhattan and kullback Leibler were seen to have the worst
807 results. However, Ex-Jaccard and Euclidean had shown a fluctuating performance. They could
808 be considered as middle-ground solutions between the best and worst measures with Ex-Jaccard
809 being superior to Euclidean.
810
811 Jaccard has been experimentally proven ineffective choice when dealing with TF-IDF-based
812 document matching. Given this fact, Jaccard was excluded from any further comparison with the
813 state-of-the-art measures. That is because Jaccard depends basically on the rate of common
814 features which is far less in the TF-IDF VSM matrix. Nevertheless, Jaccard has long been an
815 effective measure when dealing with BoW based document matching. However, Jaccard showed
816 a good clustering performance. BLAB-SM, Euclidean, SMTP, and Cosine were noted to be the
817 top performers in terms of clustering, with BLAB-SM and Euclidean being the best.
818

819 In conclusion, run time comparison showed that BLAB-SM is one of the fastest similarity
820 measures compared   to all benchmarked measures in this work, making it highly promising in
821 the machine learning and text mining fields. SMTP and Ex-Jaccard measures had been able to
822 achieve good results; yet, at the cost of their efficiency as they were slow, Ex-Jaccard being the
823 slowest measure. BLAB-SM is significantly efficient compared to SMTP while both have
824 constantly been the most effective similarity measures.

825

826 The follow-up work is planned to consider the semantic aspect along with including other
827 similarity measures [Amer & Abdalla, 2020; Oghbaie & Mohammadi Zanjireh, 2018; Sohangir
828 & Wang, 2017; Aryal, Ting, Washio, & Haffari, 2019) in the experimental study. Extensive
829 experiments are planned to be conducted on a large-scale datasets to verify the conclusions of
830 this work.

831

ACKNOWLEDGEMENT

833 The authors would like to thank and appreciate the support received from the Research Office of
834 Zayed University for providing the necessary facilities to accomplish this work.   This research
835 has been supported by Research Incentive Fund (RIF) Grant Activity Code: R19093 – Zayed
836 University, UAE.

837

REFERENCES

839 Zhang, X., & Zuo, J. (2019). Measuring similarity between sequential datasets. In ACM International Conference Proceeding Series. Association
840 for Computing Machinery. https://doi.org/10.1145/3321408.3322844
841 Gweon, H., Schonlau, M., & Steiner, S. H. A. (2019). The k conditional nearest neighbor algorithm for classification and class probability
842 estimation. PeerJ Computer Science, 2019(5). https://doi.org/10.7717/peerj-cs.194
843 Amer, A. A. (2020). On K-means clustering-based approach for DDBSs design. Journal of Big Data, 7(1). https://doi.org/10.1186/s40537-020-
844 00306-9
845 Kogan, J., Teboulle, M., & Nicholas, C. (2005). Data driven similarity measures for k-means like clustering algorithms. Information Retrieval,
846 8(2), 331–349. https://doi.org/10.1007/s10791-005-5666-8
847 Amer, A. A., Abdalla, H. I., & Nguyen, L. (2021). Enhancing recommendation systems performance using highly-effective similarity measures.
848 Knowledge-Based Systems, 106842. https://doi.org/10.1016/j.knosys.2021.106842.
849 Amer, A. A., & Abdalla, H. I. (2020). A set theory based similarity measure for text clustering and classification. Journal of Big Data, 7(1).
850 https://doi.org/10.1186/s40537-020-00344-3
851 Rachkovskij, D. A. (2017). Binary Vectors for Fast Distance and Similarity Estimation. Cybernetics and Systems Analysis, 53(1), 138–156.
852 https://doi.org/10.1007/s10559-017-9914-x
853 Gweon, H., Schonlau, M., & Steiner, S. H. B. (2019). Nearest labelset using double distances for multi-label classification. PeerJ Computer Science,
854 5. https://doi.org/10.7717/PEERJ-CS.242
855 Kanungo, T., Mount, D. M., Netanyahu, N. S., Piatko, C. D., Silverman, R., & Wu, A. Y. (2002). An efficient k-means clustering algorithms:
856 Analysis and implementation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 24(7), 881–892.
857 https://doi.org/10.1109/TPAMI.2002.1017616.
858 Holzinger, A., Schantl, J., Schroettner, M., Seifert, C., & Verspoor, K. (2014). Biomedical text mining: State-of-the-art, open problems and future
859 challenges. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in
860 Bioinformatics), 8401, 271–300. https://doi.org/10.1007/978-3-662-43968-5_16.
861 Afzali, M., & Kumar, S. (2017). Comparative Analysis of Various Similarity Measures for Finding Similarity of Two Documents. International
862 Journal of Database Theory and Application, 10(2), 23–30. https://doi.org/10.14257/ijdta.2017.10.2.02.
863 Oghbaie, M., & Mohammadi Zanjireh, M. (2018). Pairwise document similarity measure based on present term set. Journal of Big Data, 5(1).
864 https://doi.org/10.1186/s40537-018-0163-2
865 Sohangir, S., & Wang, D. (2017). Improved sqrt-cosine similarity measurement. Journal of Big Data, 4(1). https://doi.org/10.1186/s40537-017-
866 0083-6
867 Lin, Y. S., Jiang, J. Y., & Lee, S. J. (2014). A similarity measure for text classification and clustering. IEEE Transactions on Knowledge and Data
868 Engineering, 26(7), 1575–1590. https://doi.org/10.1109/TKDE.2013.19
869 Shahmirzadi, O., Lugowski, A., & Younge, K. (2019). Text similarity in vector space models: A comparative study. In Proceedings - 18th IEEE
870 International Conference on Machine Learning and Applications, ICMLA 2019 (pp. 659–666). Institute of Electrical and Electronics
871 Engineers Inc. https://doi.org/10.1109/ICMLA.2019.00120
872 Ke, W. (2017). Text retrieval based on least information measurement. In ICTIR 2017 - Proceedings of the 2017 ACM SIGIR International

Conference on the Theory of Information Retrieval (pp. 125–132). Association for Computing Machinery, Inc. https://doi.org/10.1145/3121050.3121075

White, R. W., & Jose, J. M. (2004). A study of topic similarity measures. In Proceedings of Sheffield SIGIR - Twenty-Seventh Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 520–521). Association for Computing Machinery (ACM). https://doi.org/10.1145/1008992.1009100

Lakshmi, R., & Baskar, S. (2021). Efficient text document clustering with new similarity measures. International Journal of Business Intelligence and Data Mining, 18(1), 109–126. https://doi.org/10.1504/IJBIDM.2021.111741.

Kotte, V. K., Rajavelu, S., & Rajsingh, E. B. (2020). A Similarity Function for Feature Pattern Clustering and High Dimensional Text Document Classification. Foundations of Science, 25(4), 1077–1094. https://doi.org/10.1007/s10699-019-09592

Thompson, V. U., Panchev, C., & Oakes, M. (2015). Performance evaluation of similarity measures on similar and dissimilar text retrieval. In IC3K 2015 - Proceedings of the 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (Vol. 1, pp. 577–584). SciTePress. https://doi.org/10.5220/0005619105770584

Robertson, S. (2004). Understanding inverse document frequency: On theoretical arguments for IDF. Journal of Documentation, 60(5), 503–520. https://doi.org/10.1108/00220410410560582

Zhao, R., & Mao, K. (2018). Fuzzy Bag-of-Words Model for Document Representation. IEEE Transactions on Fuzzy Systems, 26(2), 794–804. https://doi.org/10.1109/TFUZZ.2017.2690222

Joulin, A., Grave, E., Bojanowski, P., & Mikolov, T. (2017). Bag of tricks for efficient text classification. In 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017 - Proceedings of Conference (Vol. 2, pp. 427–431). Association for Computational Linguistics (ACL). https://doi.org/10.18653/v1/e17-2068

Heidarian, A., & Dinneen, M. J. (2016). A Hybrid Geometric Approach for Measuring Similarity Level among Documents and Document Clustering. In Proceedings - 2016 IEEE 2nd International Conference on Big Data Computing Service and Applications, BigDataService 2016 (pp. 142–151). Institute of Electrical and Electronics Engineers Inc. https://doi.org/10.1109/BigDataService.2016.14

Cordeiro De, Amorim, R., & Mirkin, B. (2012). Minkowski metric, feature weighting and anomalous cluster initializing in K-Means clustering. Pattern Recognition, 45(3), 1061–1075. https://doi.org/10.1016/j.patcog.2011.08.012

Kumar, V., Chhabra, J. K., & Kumar, D. (2014). Performance Evaluation of Distance Metrics in the Clustering Algorithms. INFOCOMP, 13(1), 38–51.

François, D., Wertz, V., & Verieysen, M. (2007). The concentration of fractional distances. IEEE Transactions on Knowledge and Data Engineering, 19(7), 873–886. https://doi.org/10.1109/TKDE.2007.1037

D'hondt, J., Vertommen, J., Verhaegen, P. A., Cattrysse, D., & Duflou, J. R. (2010). Pairwise-adaptive dissimilarity measure for document clustering. Information Sciences, 180(12), 2341–2358. https://doi.org/10.1016/j.ins.2010.02.021

Li, X., Yao, C., Fan, F., & Yu, X. (2017). A text similarity measurement method based on singular value decomposition and semantic relevance. Journal of Information Processing Systems, 13(4), 863–875. https://doi.org/10.3745/JIPS.02.0067

Kullback, S., & Leibler, R. A. (1951). On Information and Sufficiency. The Annals of Mathematical Statistics, 22(1), 79–86. https://doi.org/10.1214/aoms/1177729694

Subhashini, R., & Kumar, V. J. S. (2010). Evaluating the performance of similarity measures used in document clustering and information retrieval. In Proceedings - 1st International Conference on Integrated Intelligent Computing, ICIIC 2010 (pp. 27–31). https://doi.org/10.1109/ICIIC.2010.42

Li, B., & Han, L. (2013). Distance weighted cosine similarity measure for text classification. In Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) (Vol. 8206 LNCS, pp. 611–618). https://doi.org/10.1007/978-3-642-41278-3_74.

Arivarasan, A., & Karthikeyan, M. (2019). Data mining K-means document clustering using tfidf and word frequency count. International Journal of Recent Technology and Engineering, 8(2), 2542–2549. https://doi.org/10.35940/ijrte.B1718.078219

Zhao, Y., & Karypis, G. (2002). Comparison of Agglomerative and Partitional Document Clustering Algorithms. Performance Computing, 1–11.

Aslam, J. A., & Frost, M. (2003). An Information-theoretic Measure for Document Similarity. In SIGIR Forum (ACM Special Interest Group on Information Retrieval) (pp. 449–450). Association for Computing Machinery. https://doi.org/10.1145/860500.860545

Tata, S., & Patel, J. M. (2007). Estimating the selectivity of tf-idf based cosine similarity predicates. In SIGMOD Record (Vol. 36, pp. 75–80). https://doi.org/10.1145/1361348.1361351

Tanimoto T.. (1957). An elementary mathematical theory of classification and prediction. IBM, Armonk, NY, USA, Tech. Rep. IBM Int. Rep.

Haroutunian, E. (2011). Information Theory and Statistics. In International Encyclopedia of Statistical Science (pp. 666–667). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-04898-2_643

Amigó, E., Giner, F., Gonzalo, J., & Verdejo, F. (2020). On the foundations of similarity in information access. Information Retrieval Journal, 23(3), 216–254. https://doi.org/10.1007/s10791-020-09375

Kumar Nagwani, N. (2015, September 1). A Comment on "A Similarity Measure for Text Classification and Clustering." IEEE Transactions on Knowledge and Data Engineering. IEEE Computer Society. https://doi.org/10.1109/TKDE.2015.2451616

Aryal, S., Ting, K. M., Washio, T., & Haffari, G. (2019), February 9). A new simple and effective measure for bag-of-word inter-document similarity measurement. ArXiv. ArXiv

Figure 1

Figure 1 AND Gate

Figure 1 AND Gate

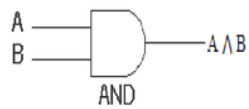



Figure 1 AND Gate

Figure 2

Figure 2 OR Gate

Figure 2 OR Gate

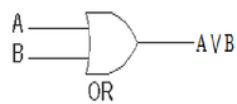



Figure 2 OR Gate

Figure 3

Figure 3 EXOR Gate

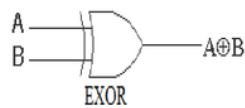



Figure 3 EXOR Gate

Figure 3 EXOR Gate

Figure 4

Figure 4. The experimental Design Flow Chart

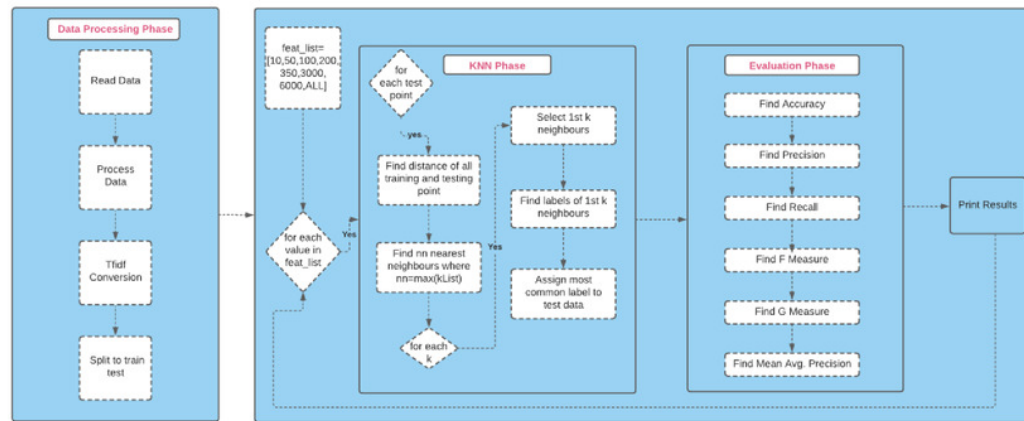



Figure 5

Figure 5 Accuracy over all Measures on all NF values - Reuters

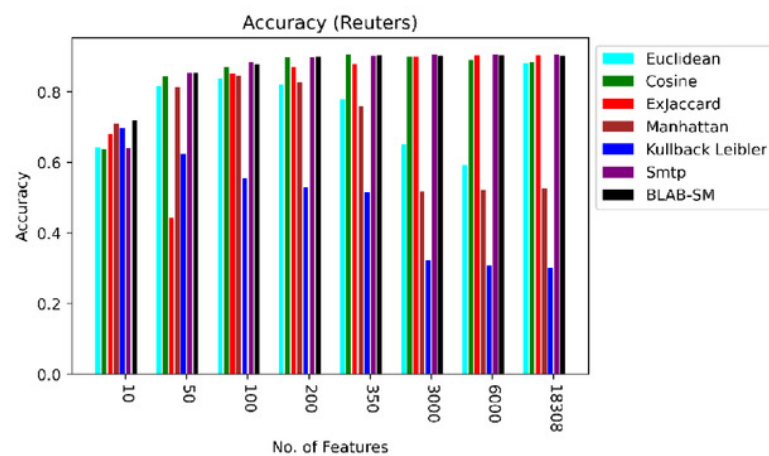
Figure 5 Accuracy over all Measures on all NF values– Average results (K=1-120; +2) –

Reuters

Figure 4 Accuracy over all Measures on all NF values– Average results (K=1-120; +2) – Reuters

Figure 6

Figure 6 F-measure over all Measures on all NF values- Reuters

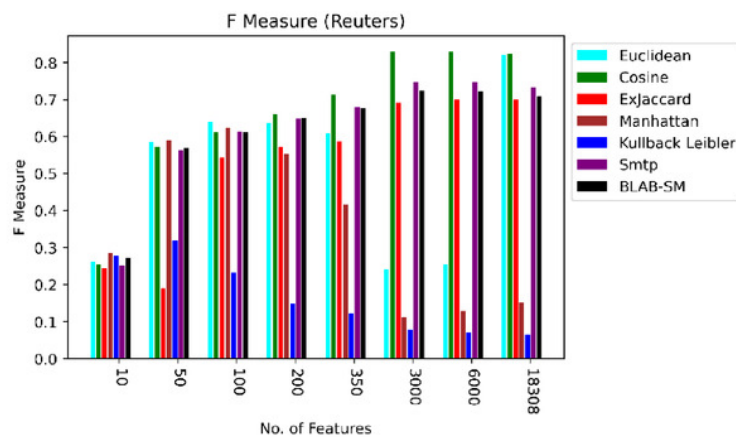
Figure 6 F-measure over all Measures on all NF values– Average results (K=1-120; +2) –

Reuters

Figure 5 F-measure over all Measures on all NF values– Average results (K=1-120; +2) – Reuters

Figure 7

Figure 7 AMP over all Measures on all NF values- Reuters

Figure 7 AMP over all Measures on all NF values– Average results (K=1-120; +2) – Reuters

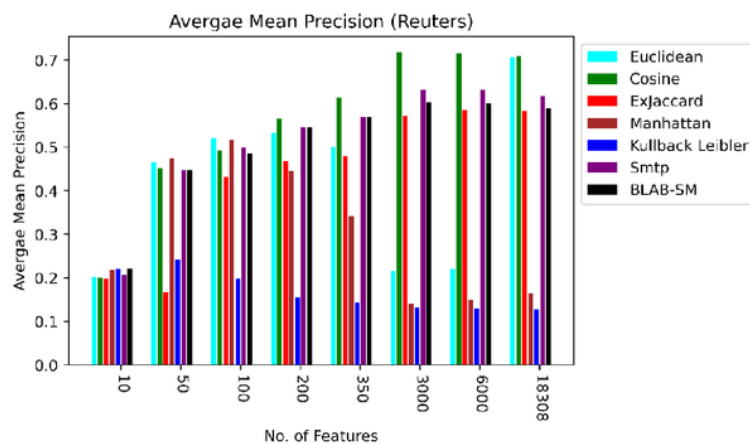



Figure 6 AMP over all Measures on all NF values– Average results (K=1-120; +2) – Reuters

Figure 8

Figure 8. Accuracy over all Measures on all NF values - Web-KB

Figure 8. Accuracy over all Measures on all NF values– Average results (K=1-120; +2) – Web-KB

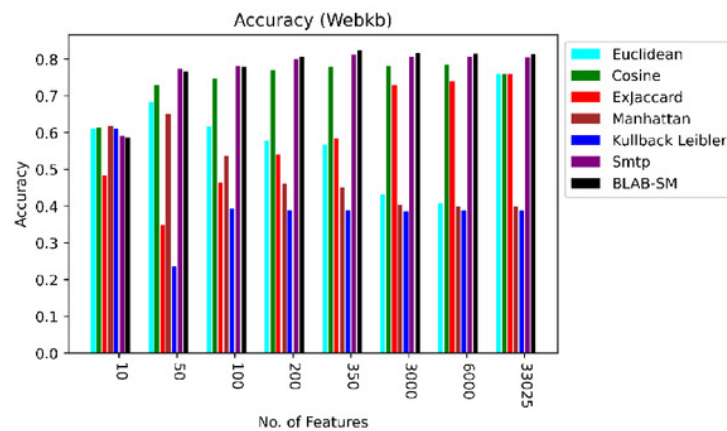



Figure 7. Accuracy over all Measures on all NF values– Average results (K=1-120; +2) – Web-KB

Figure 9

Figure 9 F-measure over all Measures on all NF values-webkb

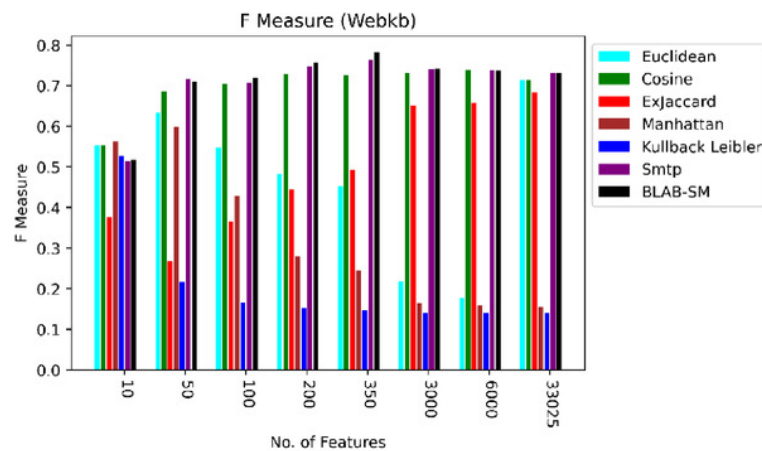
Figure 9 F-measure over all Measures on all NF values– Average results (K=1-120; +2) –

Web-KB

Figure 8 F-measure over all Measures on all NF values– Average results (K=1-120; +2) – Web-KB

Figure 10

Figure 10 AMP over all Measures on all NF values-webkb

Figure 10 AMP over all Measures on all NF values– Average results (K=1-120; +2) – Web-KB

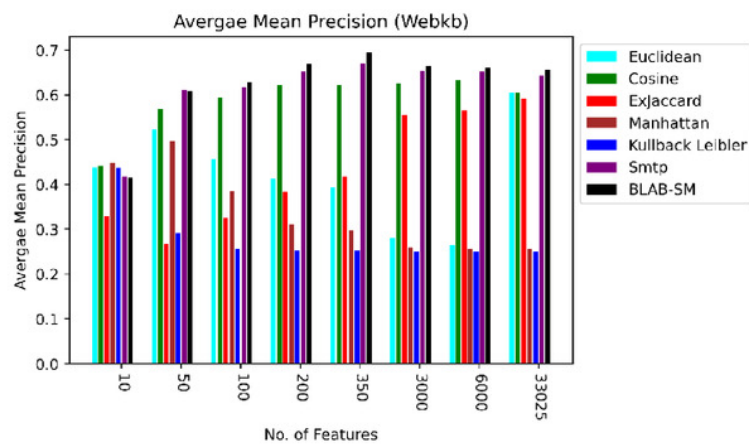



Figure 9 AMP over all Measures on all NF values– Average results (K=1-120; +2) – Web-KB

Figure 11

Figure 11. Performance of all Measures on both datasets – Reuters

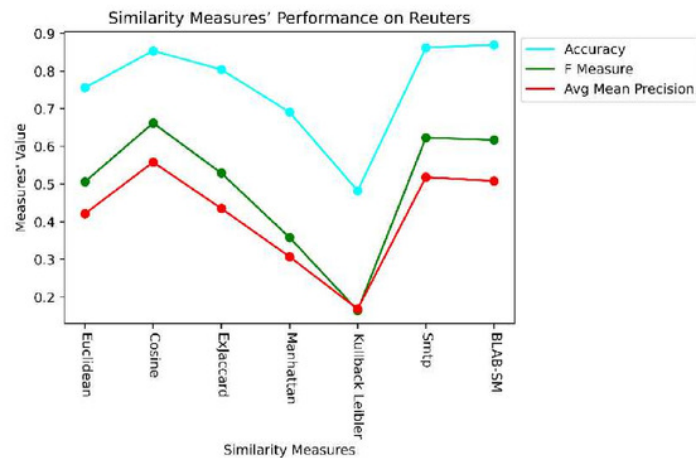



Figure 10. Performance of all Measures on both datasets – Reuters

Figure 12

Figure 12. Performance of all Measures on both datasets –Web-KB

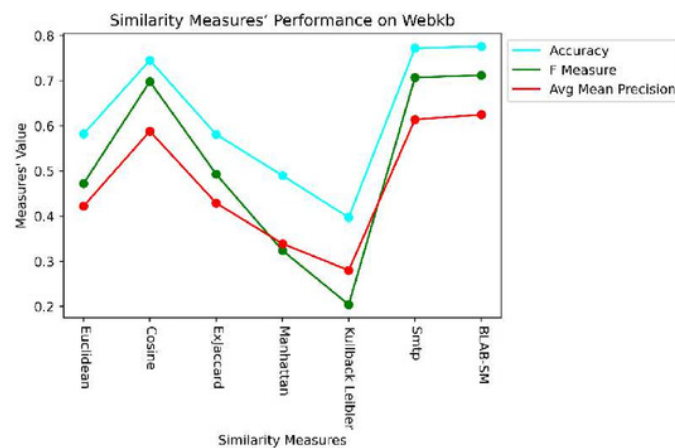



Figure 11. Performance of all Measures on both datasets –Web-KB

Figure 13

Figure 13 Execution Time – Reuters

Figure 13 Execution Time – Reuters

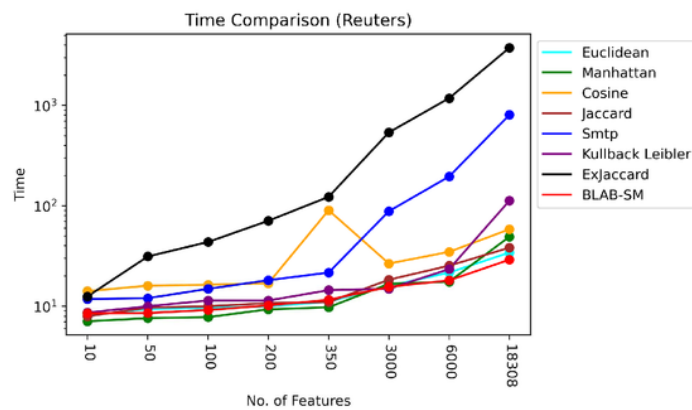



Figure 12 Execution Time – Reuters

Figure 14

Figure 14 Execution Time – Web-KB

Figure 14 Execution Time – Web-KB

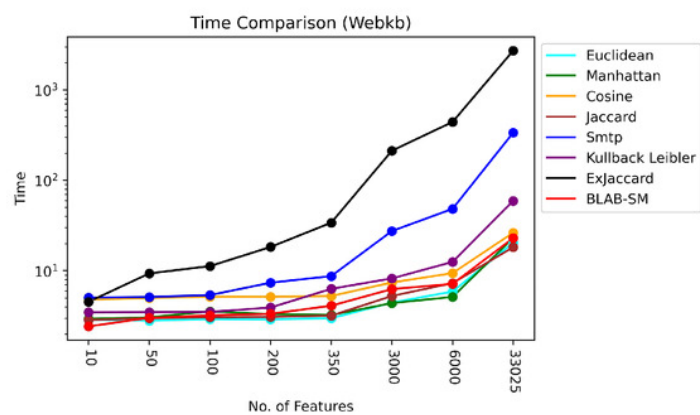



Figure 13 Execution Time – Web-KB

Table 1 A AND B table truth

Table 1 A AND B table truth

Table 1 A AND B table truth

A	B	A $\wedge$ B
0	0	0
0	1	0
1	0	
1	1	

Table 2 A OR B table truth

Table 2 A OR B table truth

Table 2 A OR B table truth

A	B	A $\lor$ B
0	0	0
0	1	1
1	0	
1	1	

Table 3 A EXOR B table truth

Table 3 A EXOR B table truth

Table 3 A EXOR B table truth

A	B	A $\oplus$ B
0	0	0
0	1	1
1	0	
1	1	

Table 4 Splitting of documents among four classes in Web-KB dataset

Table 4 Splitting of documents among four classes in Web-KB
dataset

Class	Samples
Project	504
Course	930
Student	1641
Faculty	1124
Total	4199

Table 5 Splitting of documents among eight classes in Reuters-R8 dataset

Table 5 Splitting of documents among eight classes in Reuters-R8 dataset

Class	Samples
Cq	2292
Crude	374
Earn	3923
Grain	51
Interest	271
Money-fix	293
Ship	144
Trade	326
Total	7674

Table 6. Performance Evaluation of all Measures when, NF=10 – the averaged results (K=1-120; +2)

Table 6. Performance Evaluation of all Measures when, NF=10 – the averaged results (K=1-120; +2)

Dataset	Reuters-8						Web-KB					
Similarity/ Criterion	ACC	PRE	REC	FM	GM	AMP	ACC	PRE	REC	FM	GM	AMP
Euclidean	0.642	0.345	0.272	0.262	0.504	0.203	0.611	0.598	0.540	0.553	0.679	0.438
Cosine	0.638	0.282	0.266	0.255	0.499	0.199	0.614	0.592	0.542	0.553	0.682	0.442
Ex-Jaccard	0.680	0.278	0.254	0.244	0.487	0.197	0.484	0.454	0.401	0.377	0.569	0.329
Manhattan	0.710	0.344	0.288	0.286	0.521	0.218	0.618	0.620	0.548	0.562	0.685	0.448
KLD	0.696	0.361	0.294	0.277	0.525	0.219	0.612	0.629	0.523	0.527	0.669	0.437
SMTP	0.641	0.300	0.271	0.252	0.503	0.206	0.590	0.576	0.516	0.514	0.661	0.417
BLAB-SM	0.720	0.303	0.285	0.272	0.519	0.221	0.587	0.584	0.516	0.517	0.661	0.415

Table 7. Performance Evaluation of all Measures when, NF=50– the averaged results (K=1-120; +2)

Table 7. Performance Evaluation of all Measures when, NF=50– the averaged results (K=1-120; +2)

Dataset	Reuters-8						Web-KB					
Similarity/ Criterion	ACC	PRE	REC	FM	GM	AMP	ACC	PRE	REC	FM	GM	AMP
Euclidean	0.817	0.635	0.572	0.586	0.745	0.465	0.684	0.747	0.605	0.634	0.728	0.522
Cosine	0.843	0.654	0.552	0.572	0.732	0.452	0.729	0.742	0.669	0.687	0.776	0.568
Ex-Jaccard	0.442	0.310	0.181	0.190	0.399	0.167	0.348	0.349	0.281	0.269	0.461	0.268
Manhattan	0.813	0.642	0.586	0.591	0.753	0.475	0.652	0.777	0.565	0.598	0.697	0.498
KLD	0.6245	0.602	0.282	0.320	0.505	0.242	0.236	0.527	0.316	0.217	0.496	0.292
SMTP	0.853	0.647	0.558	0.563	0.737	0.447	0.774	0.803	0.694	0.716	0.796	0.611
BLAB-SM	0.853	0.650	0.548	0.569	0.730	0.447	0.766	0.820	0.682	0.709	0.787	0.609

Table 8. Performance Evaluation of all Measures when, NF=100– the averaged results (K=1-120; +2)

Table 8. Performance Evaluation of all Measures when, NF=100– the averaged results (K=1-120; +2)

Dataset	Reuters						Web-KB					
Similarity/ Criterion	ACC	PRE	REC	FM	GM	AMP	ACC	PRE	REC	FM	GM	AMP
Euclidean	0.839	0.670	0.644	0.640	0.792	0.521	0.617	0.729	0.525	0.548	0.666	0.456
Cosine	0.870	0.686	0.592	0.610	0.760	0.493	0.747	0.776	0.685	0.705	0.788	0.593
Ex-Jaccard	0.852	0.632	0.523	0.544	0.713	0.433	0.464	0.555	0.372	0.365	0.544	0.325
Manhattan	0.847	0.686	0.638	0.624	0.788	0.517	0.536	0.772	0.426	0.429	0.587	0.386
KLD	0.555	0.659	0.211	0.232	0.432	0.197	0.394	0.386	0.260	0.166	0.442	0.256
SMTP	0.884	0.667	0.603	0.614	0.768	0.499	0.781	0.815	0.692	0.708	0.796	0.617
BLAB-SM	0.879	0.669	0.585	0.610	0.756	0.486	0.778	0.853	0.693	0.720	0.795	0.629

Table 9. Performance Evaluation of all Measures when, NF=200– the averaged results (K=1-120; +2)

Table 9. Performance Evaluation of all Measures when, NF=200– the averaged results (K=1-120; +2)

Dataset	Reuters						Web-KB					
Similarity/ Criterion	ACC	PRE	REC	FM	GM	AMP	ACC	PRE	REC	FM	GM	AMP
Euclidean	0.821	0.738	0.607	0.637	0.766	0.533	0.578	0.742	0.470	0.482	0.624	0.414
Cosine	0.897	0.739	0.642	0.661	0.793	0.565	0.770	0.797	0.710	0.729	0.806	0.621
Ex-Jaccard	0.871	0.644	0.554	0.572	0.735	0.468	0.540	0.638	0.443	0.446	0.604	0.383
Manhattan	0.828	0.733	0.510	0.553	0.700	0.446	0.461	0.792	0.327	0.280	0.504	0.311
KLD	0.530	0.611	0.161	0.149	0.375	0.155	0.389	0.350	0.254	0.152	0.437	0.253
SMTP	0.897	0.721	0.635	0.648	0.789	0.545	0.800	0.839	0.725	0.748	0.818	0.651
BLAB-SM	0.900	0.719	0.629	0.650	0.785	0.545	0.807	0.874	0.727	0.757	0.820	0.669

Table 10 Performance Evaluation of all Measures when, NF=350– the averaged results (K=1-120; +2)

Table 10 Performance Evaluation of all Measures when, NF=350– the averaged results (K=1-120; +2)

Dataset	Reuters						Web-KB					
Similarity/ Criterion	ACC	PRE	REC	FM	GM	AMP	ACC	PRE	REC	FM	GM	AMP
Euclidean	0.780	0.785	0.551	0.608	0.727	0.501	0.567	0.737	0.451	0.452	0.610	0.394
Cosine	0.905	0.772	0.686	0.714	0.820	0.613	0.778	0.797	0.711	0.726	0.809	0.621
Ex-Jaccard	0.880	0.732	0.571	0.588	0.747	0.479	0.583	0.674	0.486	0.492	0.639	0.417
Manhattan	0.760	0.755	0.368	0.417	0.583	0.341	0.451	0.618	0.310	0.245	0.490	0.297
KLD	0.516	0.548	0.146	0.123	0.357	0.142	0.388	0.368	0.252	0.147	0.435	0.252
SMTP	0.902	0.778	0.657	0.679	0.803	0.570	0.812	0.852	0.741	0.765	0.830	0.670
BLAB-SM	0.904	0.770	0.647	0.677	0.800	0.570	0.825	0.881	0.753	0.783	0.838	0.695

Table 11 Performance Evaluation of all Measures when, NF=3000 the averaged results (K=1-120; +2)

Table 11 Performance Evaluation of all Measures when, NF=3000 the averaged results

(K=1-120; +2)

Table 11 Performance Evaluation of all Measures when, NF=3000– the averaged results (K=1-120; +2)

Dataset	Reuters						Web-KB					
Similarity/ Criterion	ACC	PRE	REC	FM	GM	AMP	ACC	PRE	REC	FM	GM	AMP
Euclidean	0.652	0.576	0.228	0.242	0.451	0.216	0.431	0.627	0.293	0.218	0.473	0.281
Cosine	0.901	0.904	0.785	0.830	0.876	0.718	0.782	0.797	0.717	0.730	0.813	0.626
Ex-Jaccard	0.900	0.874	0.657	0.691	0.803	0.571	0.729	0.776	0.641	0.650	0.759	0.555
Manhattan	0.517	0.224	0.141	0.112	0.350	0.140	0.404	0.286	0.263	0.165	0.446	0.260
KLD	0.323	0.245	0.135	0.077	0.345	0.132	0.386	0.174	0.250	0.141	0.433	0.250
SMTP	0.907	0.895	0.703	0.748	0.830	0.632	0.807	0.854	0.724	0.741	0.819	0.653
BLAB-SM	0.902	0.889	0.670	0.725	0.809	0.603	0.816	0.876	0.726	0.742	0.821	0.665

Table 12. Performance Evaluation of all Measures when, NF=6000 the averaged results
(K=1-120; +2)

Table 12. Performance Evaluation of all Measures when, NF=6000– the averaged results (K=1-120; +2)

Dataset	Reuters						Web-KB					
Similarity/ Criterion	ACC	PRE	REC	FM	GM	AMP	ACC	PRE	REC	FM	GM	AMP
Euclidean	0.593	0.714	0.227	0.253	0.446	0.220	0.407	0.448	0.269	0.177	0.451	0.265
Cosine	0.891	0.900	0.786	0.829	0.876	0.716	0.785	0.795	0.728	0.739	0.819	0.632
Ex-Jaccard	0.904	0.882	0.667	0.700	0.809	0.584	0.739	0.787	0.652	0.658	0.767	0.565
Manhattan	0.521	0.327	0.151	0.129	0.362	0.150	0.398	0.263	0.260	0.159	0.443	0.257
KLD	0.308	0.254	0.131	0.069	0.339	0.130	0.387	0.176	0.251	0.140	0.434	0.250
SMTP	0.907	0.897	0.702	0.748	0.829	0.632	0.806	0.853	0.722	0.738	0.818	0.651
BLAB-SM	0.904	0.886	0.668	0.722	0.808	0.600	0.815	0.871	0.724	0.737	0.820	0.661

Table 13. Performance Evaluation of all Measures when, NF= the whole size (Reuters =18308, web-kb=33025 features) – the averaged results (K=1-120; +2)

Table 13. Performance Evaluation of all Measures when, NF= the whole size (Reuters =18308, web-kb=33025 features) – the averaged results (K=1-120; +2)

Table 13. Performance Evaluation of all Measures when, NF= the whole size (Reuters =18308, web-kb=33025 features) – the averaged results (K=1-120; +2)

Dataset	Reuters						Web-KB					
Similarity/ Criterion	ACC	PRE	REC	FM	GM	AMP	ACC	PRE	REC	FM	GM	AMP
Euclidean	0.882	0.895	0.779	0.821	0.870	0.706	0.760	0.773	0.712	0.714	0.808	0.604
Cosine	0.884	0.898	0.779	0.823	0.871	0.709	0.760	0.773	0.712	0.714	0.808	0.604
Ex-Jaccard	0.904	0.883	0.666	0.700	0.809	0.583	0.760	0.800	0.679	0.685	0.787	0.591
Manhattan	0.527	0.378	0.167	0.151	0.379	0.164	0.398	0.223	0.258	0.155	0.441	0.256
KLD	0.301	0.255	0.129	0.065	0.337	0.128	0.387	0.177	0.251	0.140	0.434	0.250
SMTP	0.905	0.894	0.688	0.733	0.821	0.617	0.804	0.851	0.718	0.730	0.816	0.644
BLAB-SM	0.902	0.848	0.659	0.708	0.803	0.590	0.814	0.865	0.723	0.731	0.820	0.656

Table 14. Performance Evaluation of all Measures when the average of averaged results is considered

Table 14. Performance Evaluation of all Measures when the average of averaged results is considered (Reuters =18308, web-kb=33025 features) – the averaged of averaged results (NF=10-All features)

Dataset	Reuters						Web-KB					
Similarity/ Criterion	ACC	PRE	REC	FM	GM	AMP	ACC	PRE	REC	FM	GM	AMP
Euclidean	0.756	0.670	0.485	0.506	0.663	0.421	0.582	0.675	0.483	0.472	0.630	0.422
Cosine	0.854	0.729	0.636	0.662	0.778	0.558	0.745	0.759	0.684	0.698	0.788	0.588
Ex-Jaccard	0.804	0.654	0.509	0.529	0.688	0.435	0.581	0.629	0.494	0.493	0.641	0.429
Manhattan	0.691	0.511	0.356	0.358	0.554	0.307	0.490	0.544	0.370	0.324	0.536	0.339
KLD	0.482	0.442	0.186	0.164	0.402	0.168	0.397	0.348	0.295	0.204	0.472	0.280
SMTP	0.862	0.725	0.602	0.623	0.760	0.518	0.772	0.805	0.692	0.707	0.794	0.614
BLAB-SM	0.870	0.717	0.586	0.617	0.751	0.508	0.776	0.828	0.693	0.712	0.795	0.625

Table 15. Purity (mostly known as "Accuracy") - K-means performance

Table 15. Purity (mostly known as "Accuracy") - K-means performance

Similarity measure/ Metric	K =5		K = 10		K= Number of classes	
	Reuters – 18308 features	Web-KB - 33025 features	Reuters – 18308 features	Web-KB - 33025 features	Reuters – 18308 features (K=4)	Web-KB - 33025 features (K=8)
Euclidean	0.6348979	0.6701596	0.7467169	0.6244344	0.6376284	0.6480114
Cosine	0.6200751	0.6546797	0.6641529	0.6263396	0.6727344	0.6225292
Jaccard	0.55571447	0.6468207	0.6736445	0.6663491	0.6222858	0.6320553
Ex-Jaccard	0.5942478	0.6434865	0.6970484	0.6261014	0.6358080	0.6618242
KLD	0.6091535	0.6170517	0.7253933	0.6477733	0.6618125	0.6525363
Manhattan	0.5648979	0.5901565	0.6368767	0.5845236	0.6076209	0.6281809
BLAB-SM	0.6313405	0.6037151	0.6965284	0.6954036	0.6459498	0.63824720
SMTP	0.6261864	0.6408668	0.6885970	0.6356275	0.6374983	0.63372231

Table 16. Entropy - K-means performance

Table 16. Entropy - K-means performance

Similarity measure/ Metric	K =5		K = 10		K= Number of classes	
	Reuters – 18308 features	Web-KB - 33025 features	Reuters – 18308 features	Web-KB - 33025 features	Reuters – 18308 features (K=4)	Web-KB - 33025 features (K=8)
Euclidean	0.4181014	0.6083767	0.3324207	0.6338744	0.4066976	0.6489474
Cosine	0.5079926	0.6226576	0.3728674	0.6445081	0.3673066	0.6287172
Jaccard	0.4551138	0.6505793	0.4093773	0.6063160	0.4044194	0.6138461
Ex-Jaccard	0.4210397	0.6307335	0.3478137	0.6219528	0.4078592	0.6325090
KLD	0.5174439	0.6295514	0.3337954	0.6137706	0.3724656	0.6335062
Manhattan	0.4481032	0.6387609	0.3225643	0.6034765	0.3866542	0.6089534
BLAB-SM	0.4941582	0.6273114	0.3253195	0.5561895	0.4045278	0.6356724
SMTP	0.4647862	0.6486512	0.3147964	0.5373072	0.4038267	0.6341636

Table 17. Measure Stability Status

Table 17. Measure Stability Status

NF/Measure	Euclidean		Cosine		Ex-Jaccard		Manhattan		kullback Leibler		SMTP		BLAB-SM	
	R	W	R	W	R	W	R	W	R	W	R	W	R	W
10		3		2			3	3	3	1			3	
50	2		1	3			2				1	3	1	3
100	2			3			2				3	3	1	3
200			3	3							3	3	3	3
350			3	3							3	3	3	3
3000			3	3							3	3	3	3
6000			2	3	1						3	3	3	3
All features	2	3	2	3	1						3	3	1	3
Sum	6	6	14	23	2		7	3	3	1	19	21	18	21
Stable points	12		39		4		11		4		40		39	

Table 18. Reuters - Statistical Significance of the experimental results of BLAB-SM against all measures

Table 18. Reuters - Statistical Significance of the experimental results of BLAB-SM against all measures

Measure	MAE	MSE	Std Error	Confidence Interval
Manhattan	0.473236691	0.22496147	0.0040998006	0.51855962- 0.53496698
Euclidean	0.117193134	0.01375278	0.0005560327	0.88169424- 0.88391948
Cosine	0.114459829	0.01312222	0.0005939627	0.88435165- 0.88672868
Jaccard	0.497886917	0.25102953	0.0072320509	0.48764178- 0.51658438
KLD	0.496776287	0.24930169	0.0064743325	0.49026860- 0.51617882
SMTP	**0.095485359**	**0.00931860**	0.0018312414	**0.90085033- 0.90817894**
Ex-Jaccard	**0.091511611**	**0.00845275**	0.0011429639	**0.90620132- 0.91077545**
BLAB-SM	**0.099111785**	**0.01006081**	0.0019902300	**0.89690577- 0.90487065**

Table 19. Reuters -Statistical significance with non-parametric Wilcoxon for BLAB-SM against all measures

Table 19. Reuters -Statistical significance with non-parametric Wilcoxon for BLAB-SM against all measures

Measure	Wilcoxon Z Value	Wilcoxon P Value
Manhattan	-6.7359808238	0.0
Euclidean	-6.4488297578	0.00000000001
Cosine	-5.6390451991	0.0000000171
Jaccard	-6.7365741193	0.0
KLD	-6.7365512974	0.0
SMTP	-5.9849832403	0.0000000022
Ex-Jaccard	-4.9985622992	0.0000005776
BLAB-SM	Nan	nan

Table 20. Reuters -Statistical significance with paired t-test (DF is degrees of freedom) for BLAB-SM against all measures

Table 20. Reuters -Statistical significance with paired t-test (DF is degrees of freedom) for BLAB-SM against all measures

Table 20. Reuters -Statistical significance with paired t-test (DF is degrees of freedom) for BLAB-SM against all measures

Measure	t-score	tTest p-value	DF
Manhattan	-124.5610100138	0.0	59.0
Euclidean	-8.2792339235	0.0	59.0
Cosine	-6.8900087347	0.0000000042	59.0
Jaccard	-49.7455369193	0.0	59.0
KLD	-54.7808677782	0.0	59.0
SMTP	10.3197557538	0.0	59.0
Ex-Jaccard	5.6215858102	0.000000543	59.0
BLAB-SM	Nan	nan	59.0

Table 21. Web-KB - Significance Test of BLAB-SM against all measures

Table 21. Web-KB - Significance Test of BLAB-SM against all measures

Measure	MAE	MSE	Std Error	Confidence Interval
Manhattan	0.602047015	**0.362962271**	0.002891546	0.39216701 - 0.40373895
Euclidean	0.239183835	0.058129364	0.003916753	0.75297875 - 0.76865356
Cosine	0.238087691	0.057424802	0.003509637	0.75488953 - 0.76893507
Jaccard	0.776096143	0.602833150	0.002909542	0.21808187 - 0.22972583
KLD	0.777548864	0.605136971	0.003040655	0.21636679- 0.22853547
SMTP	**0.195668251**	**0.038413104**	**0.001455103**	**0.80142009 - 0.80724340**
Ex-Jaccard	0.240649762	0.058737734	0.003709056	0.75192843 - 0.76677204
BLAB-SM	**0.186283068**	**0.034801849**	**0.001294012**	**0.81112761 - 0.81630624**

Table 22. Web-KB - Statistical significance with non-parametric Wilcoxon for BLAB-SM against all measures

Table 22. Web-KB - Statistical significance with non-parametric Wilcoxon for BLAB-SM against all measures

Measure	Wilcoxon Z Value	Wilcoxon P Value
Manhattan	-6.7365284757	0.0
Euclidean	-6.7359351922	0.0
Cosine	-6.7358895615	0.0
Jaccard	-6.7368936511	0.0
KLD	-6.7368936511	0.0
SMTP	-6.213256068	0.0000000005
Ex-Jaccard	-6.7358667465	0.0
BLAB-SM	nan	nan

Table 23. Web-KB - Statistical significance with paired t-test (DF stands for degrees of freedom)

Table 23. Web-KB - Statistical significance with paired t-test (DF stands for degrees of freedom)

Measure	t-score	tTest p-value	DF
Manhattan	-116.7083551117	0.0	59.0
Euclidean	-13.9243714309	0.0	59.0
Cosine	-15.2824025577	0.0	59.0
Jaccard	-158.01549932	0.0	59.0
KLD	-153.0210104997	0.0	59.0
SMTP	-11.223396507	0.0	59.0
Ex-Jaccard	-14.5841724659	0.0	59.0
BLAB-SM	nan	nan	59.0

Table 24. Clustering Process - Points and Rank of Similarity Measures

Table 18. Clustering Process - Points and Rank of Similarity Measures

Similarity Measure	Points (out of 12)	Rank
Euclidean	6	1
Cosine	5	2
Jaccard	3	-
Ex-Jaccard	3	-
KLD	5	2
Manhattan	4	3
BLAB-SM	6	1
SMTP	4	2

Table 25. Comparison of Classification Run Time (in Minutes) of BLAB-SM against all Measures on both Datasets

4   Table 19. Comparison of Classification Run Time (in Minutes) of BLAB-SM against all Measures on both Datasets

Measure/ Dataset	Reuters (18308 features)	Improvement Rate	Web-KB (33025 features)	Improvement Rate
BLAB-SM	29	-	23	-
Cosine	58.28	50.24%	26.37	12.78%
SMTP	805	96.40%	337	93.18%
Euclidean	34.17	15.13%	20.42	-11.22%
Manhattan	49.14	40.98%	23.10	0.44%
Jaccard	38.19	24.06%	18.24	-20.69%
kullback Leibler	112.35	74.19%	59.37	61.02%