Menoufia Univeristy
Faculty of Electronic Engineering in Menouf
Computer Science and Engineering Department

Evaluation and Enhancement of Arabic Text Retrieval Models

تقييم وتحسين آداء نماذج استرجاع النصوص العربية

Presented by:

Eng. Ayat Elnabawey Elnahas

Supervised by :

Prof. Mohamed Nour Elsayed

Professor of Computer Engineering

Former Vice-President of Electronics Research Institute, Cairo

Prof. Nawal Ahmed El-Fishawy

Professor of Computer Scince and Engineering

Faculty of Electronic Engineering, Menoufia Univeristy

Dr. Maha Saad El-Deen Tolba

Lecturer of Computer Science and Engineering

Faculty of Electronic Engineering, Menoufia Univeristy

Agenda

- **Problem Definition**
- **Introduction to Information Retrieval System**
- **Adopted Arabic Information Retrieval Model**
  - **Document Collection Preprocessing**
  - **Indexing**
  - **User Query**
  - **Matching, Ranking and Retrieval**
- **Proposed Approaches**
  - **Query Expansion Using Relevance Feedback**
  - **Query Expansion Using Semantics of Keywords**
  - **Hybrid of Relevance Feedback and Semantics Approach**

Agenda (Cont.)

- **Implementation and Experimental Results**
- **Introduction to Query Expansion Using Word Embedding**
- **Continuous Bag-of-Words (CBOW) model**
- **Skip-Gram model**
- **Analysis of a Word Embedding Approach**
- **Implementation of Hybrid Method**
- **Experimental Results**

Agenda (Cont.)

- **Text Classification**
  - **The Process of Arabic Text Classification**
- **The Adopted Classification Algorithms**
  - **Decision Tree**
  - **Naïve Bayes**
  - **Support Vector Machine**
- **Implementation and Experimental Results of Three Classifiers**

Agenda (Cont.)

- **Feature Selection Methods**
  - **Term Weighting (TF-IDF)**
  - **Chi-square**
  - **Gini Index (GI)**
  - **Information Gain (IG)**
  - **Semantic Fusion**
- **Experimental Results**
- **A Chosen Deep Learning Model**
- **Convolution Neural Networks**
- **Experimental Results**
- **Conclusion and Future Work**

❑ Problem Definition

- Due to the huge amount of information uploaded on the Internet, the process of finding information on a specific topic is becoming very important.

- This work aims to

- Match and retrieve the most relevant documents to the user query.
- Improve the performance of retrieval process
- Classify the retrieved documents

❑ Problem Definition (cont.)

- Several models for text retrieval were developed, but most of Arabic information retrieval models do not satisfy the user needs.

- The Arabic language is more powerful and has special features .

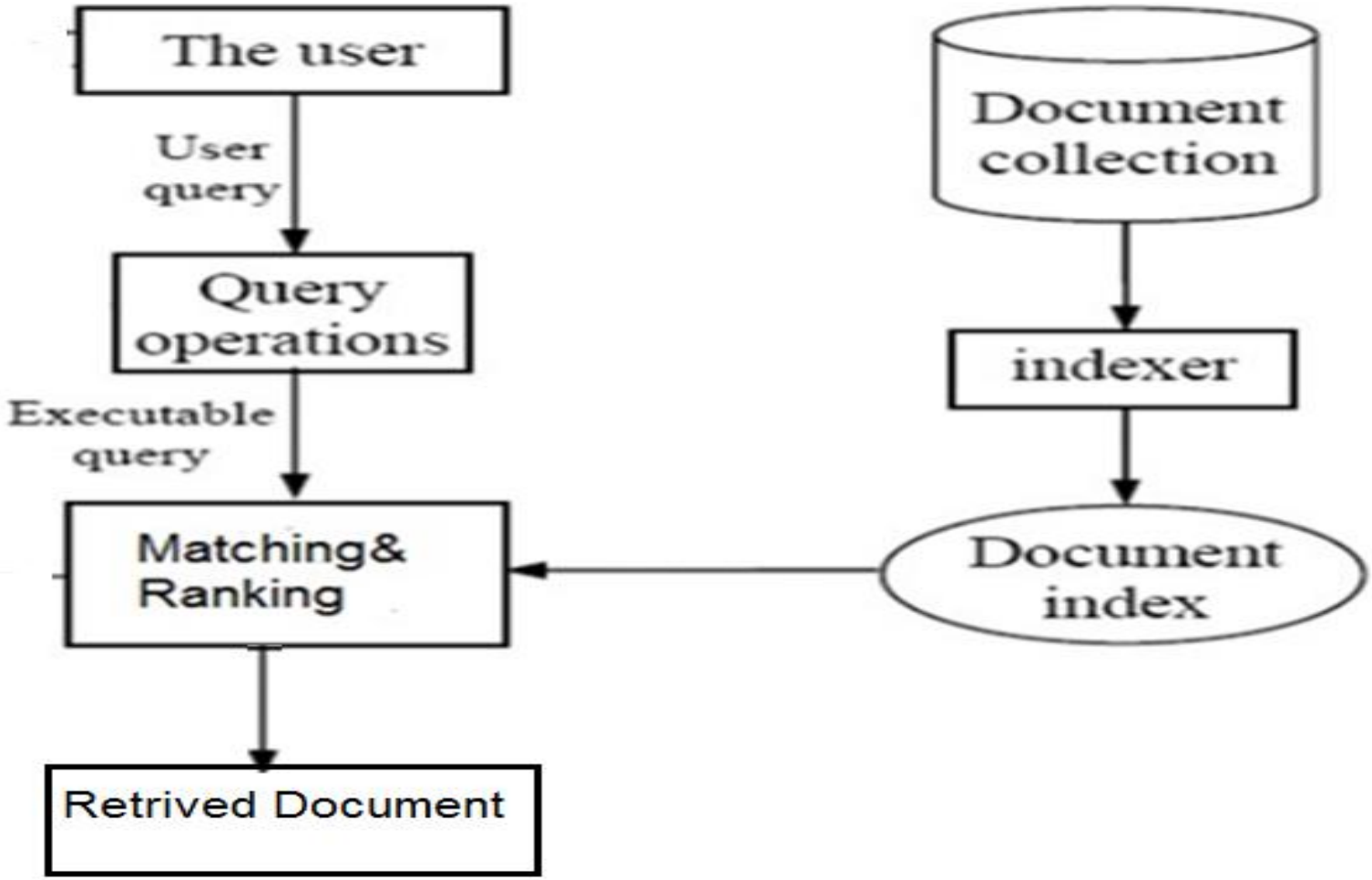
❑ Introduction to Information Retrieval System

- A good information retrieval system should retrieve only those documents that satisfy the user needs.

- There are a lot of models used in information retrieval systems. This includes but not limited to: Boolean model, Probabilistic model, Vector Space model, and others.

- An information retrieval system contains several modules mainly: document collection, query processing, matching operations and query performance.

❑ Adopted Arabic Information Retrieval Model



The Main Modules of an Information Retrieval System

❑ Adopted Arabic Information Retrieval Model (cont.)

➢ Preprocessing Operations

- The preprocessing steps are done on the document terms before building the index and on the user query before matching process.
- The preprocessing steps involve:

Tokenization

Removal of stop-words

Stemming

❑ Adopted Arabic Information Retrieval Model (cont.)

▪ Tokenization

Tokenization; means splitting text into tokens.

A document title before tokenization

أهم المركبات المسموح بها في الزراعة العضوية لمقاومة الأمراض والحشرات

A document title after tokenization

أهم,المركبات,المسموح,بها,في,الزراعة,العضوية, لمقاومة,الأمراض,و,الحشرات

❑ Adopted Arabic Information Retrieval Model (cont.)

▪ Removal of Stop-words

Removal of stop words means rejecting the useless words like preposition, pronoun, specifiers, modifiers, and other tools. Examples of the stop words are: - **من، هو، هي، الذي إلي، علي، في، .**

A document title before removal of stop-words

أهم,المركبات,المسموح,بها,في,الزراعة,العضوية, لمقاومة,الأمراض,و,الحشرات

A document title after removal of stop-words

أهم، المركبات، المسموح، الزراعه، العضويه، مقاومة، الأمراض، الحشرات

❑ Adopted Arabic Information Retrieval Model (cont.)

- **Stemming**

  Stemming aims at reducing all the derivational variants of words into a common form called the stem.

A document title before stemming

أهم، المركبات، المسموح، الزراعه، العضويه، مقاومة، الأمراض، الحشرات

A document title after stemming

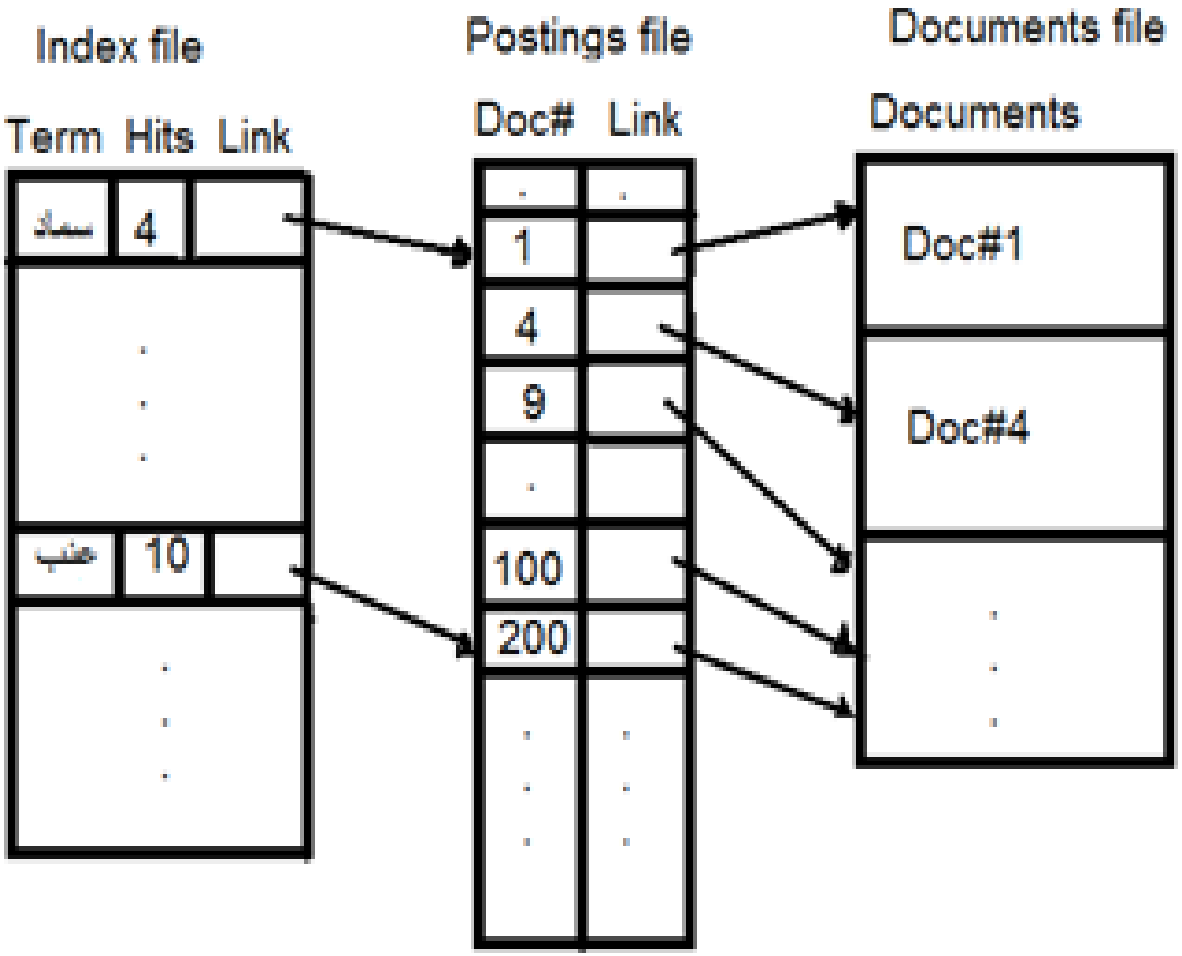
أهم، مركب، مسموح، زراع، عضوي، مقاوم، أمراض، حشر

❑ Adopted Arabic Information Retrieval Model (cont.)

➢ Indexing

- Indexing is the process of choosing a term or a number of terms that can represent what the document contains.
- Each document is represented by a set of important terms.
- Terms are weighted and stored in an index (as index terms) without any repetition.
- The index contains document number, terms, frequency /weight and other useful information such as the number of documents that contain each term.

❑ Adopted Arabic Information Retrieval Model (cont.)



Adopted Arabic Information Retrieval Model (cont.)

User Query

- The querying stage is handled exactly like the document.
- The preprocessing steps: tokenization, removal of stop-words, and stemming are done on the input user query.
- The user query may be phrase, or sentence containing a set of keywords.

User Query	No. of Keywords
زراعة الخضروات	2
صادرات مصر من القمح	3
أهمية البلح وطرق تجفيفة	4

User Query	No. of Keywords
محصول العنب	2
الجديد فى محصول العنب	3
الجديد فى إنتاج محصول العنب	4

❑ Adopted Arabic Information Retrieval Model (cont.)

➢ Matching, Ranking and Retrieval

- The matching process is done between the query keywords and document terms.
- The Vector Space Model (VSM) is used to represent both the documents and the query.
- The VSM is an algebraic model where it uses non-binary weights that are assigned to the index terms of documents and queries

❑ Adopted Arabic Information Retrieval Model (cont.)

➢ Matching, Ranking and Retrieval (cont.)

- Documents can be retrieved and ranked by matching the query vector versus the document vector to compute the score or similarity.

- The retrieved documents are ranked according to the similarity to the user query .

Adopted Arabic Information Retrieval Model (cont.)

Matching, Ranking and Retrieval (cont.)

$$sim\left(d_j q_i\right) = \frac{\sum_{i=1}^{t} w_{ij} * Wiq}{\sqrt{\sum_{i=1}^{t} w_{ij}^2} * \sqrt{\sum_{i=1}^{t} w_{iq}^2}} \qquad (1)$$

where

$$w_{ij} = tf_{ij*}\ idf_i$$

tf_{ij} is the number of occurrence of term i in the document j

$$idf_j = log_2 \frac{N}{n_i}$$

❑ First Approach

➢ Query Expansion using Relevance Feedback

- Query expansion means adding extra new terms to the keywords of the initial query.
- In this approach the query is modified by selecting relevant textual keywords for expanding the query and avoid the non-related textual words.
- The process of query expansion based on the principle of user relevance feedback is described as follows:
- The original keywords of the user query, after doing the preprocessing operations, are matched against the index terms.

❑ First Approach(cont.)

➢ Query Expansion using Relevance Feedback (cont.)

- The retrieved documents are presented from the highest to lowest values depending on the similarity values.
- The terms' descriptors for only those retrieved documents with similarity values $\geq max_{th}$ are chosen to be added to the original query keywords.
- Let S_1 be the set that collects all relevant retrieved documents that satisfy the threshold condition max_{th}.
- $S_1 = \{ d_1, d_2, \ldots\ldots\ldots max_{th}\}$

First Approach (cont.)

Query Expansion using Relevance Feedback (cont.)

- A minimum threshold value (min_{th}) of documents similarities is defined.
- The terms' descriptors for only those retrieved documents with similarity values $\leq min_{th}$ will be eliminated from the query
- Let S_2 be the set that gathers all non-relevant retrieved documents and the min_{th} condition is satisfied. $S_2 = \{ d_1, d_2, \ldots\ldots\ldots d_y\}$
- The query can be expanded by adding the terms of the selected relevant documents from S1 and also eliminating those terms of the chosen irrelevant documents from S2.
- $q_{exp} = q_{user} + \sum_{di \in s_1} d_i - \sum_{dj \in s_2} d_j$ (2)

❑ Implementation and Experimental Results

➢ Simulation Environment

- Software (JAVA programming language besides Lucene APIS.)

- Hardware (HP-Labtop with a processor 2.5 GHZ, and Windows-7 operating systems.)

❑ Implementation and Experimental Results (cont.)

➢ Document Collection Dataset

- The documents in the dataset are acquired from different Arabic websites. http://www.kenanaonline.net/page/Agriculture and http://www.zeraiah.net/index.php/baydar.
- The test-bed documents are in the agriculture field.
- The test-bed contains 1400 documents.
- Each document has a document title and contents.

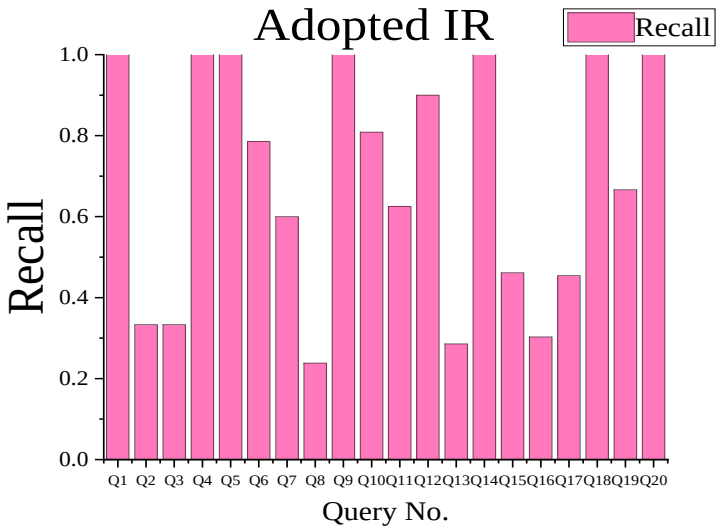
❑ Implementation and Experimental Results (cont.)

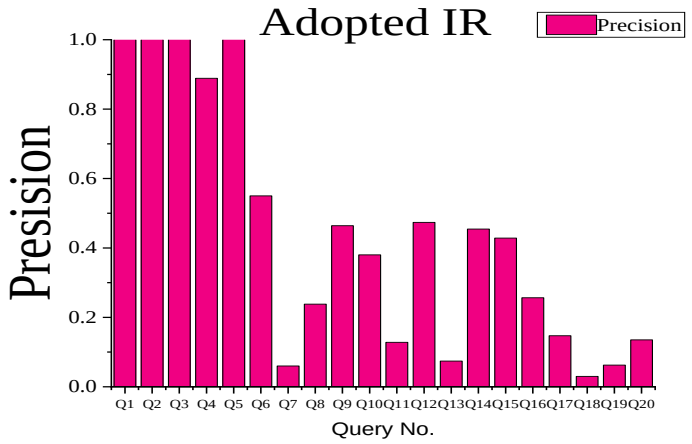
➢ Performance metrics

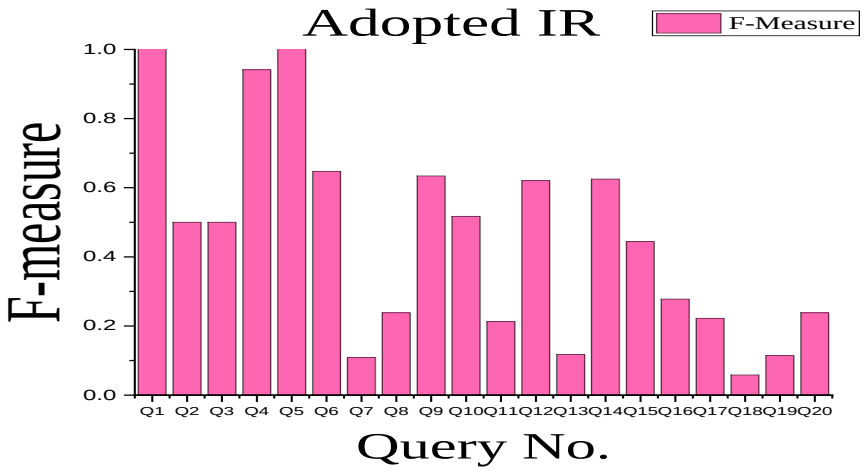
- **Recall** $= \frac{number\ of\ relevant\ retrieved\ documents}{number\ of\ retrieved\ documents}$
- **Precision** $= \frac{number\ of\ relevant\ retrieved\ documents}{number\ of\ relevant\ documents}$
- **F-measure** $= \frac{2(Recall * Precision)}{(Recall + Precision)}$

❑ Implementation and Experimental Results (cont.)

➢ Adopted IR Model

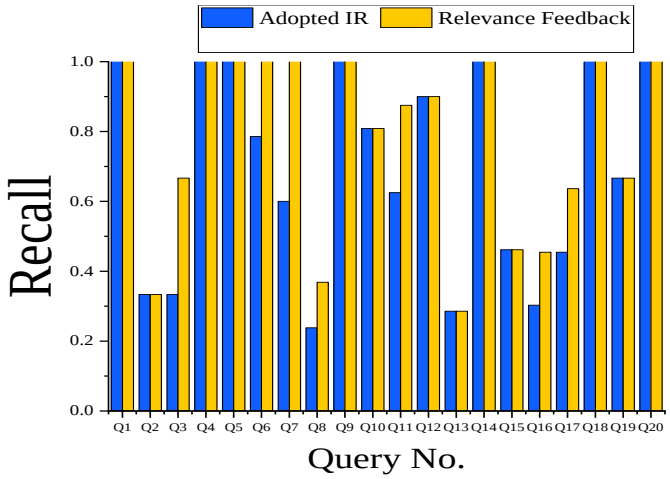


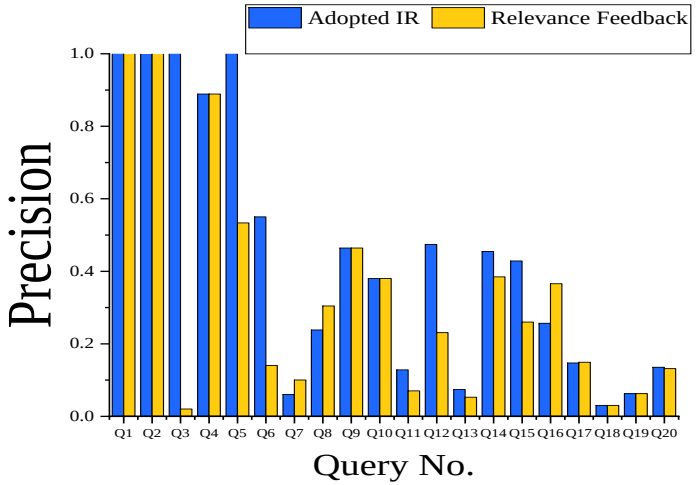


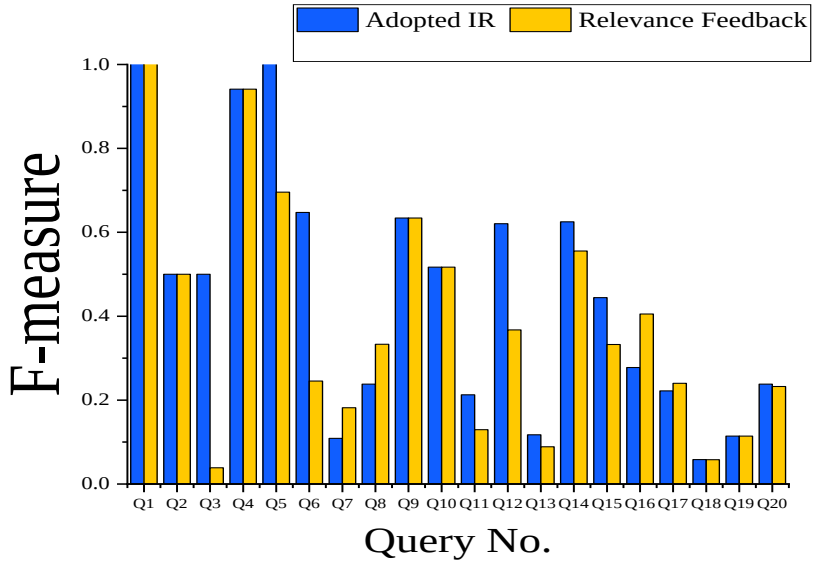


❑ Implementation and Experimental Results (cont.)

➢ Adopted IR and Relevance Feedback







❑ Second Approach

➢ Query Expansion using Semantics of Keywords

Expanding the query to include more or extra keywords will improve the performance of the retrieval model as it presents more relevant documents to the user.

- $Q = \{q_1, q_2, q_3, \ldots..q_r\}$
- $q_r = \{k_1, k_2, \ldots.k_m\}$
- $S_{ki} = \{S_{i1}, S_{i2}, S_{i3}, \ldots, S_{in}\}$
- $S(q_r) = \{S_{k1}, S_{k2}, \ldots..S_{km}\}$
- $q_{exp} = q_{user} + S(q_r)$ (3)

where
Q: is a set of queries entered separately from the user.
q_r: represent each query
k_m: query keywords
S_{ki}: synonyms associated to a keyword ki
$S(_{qr})$: set of lists

Second Approach(cont.)

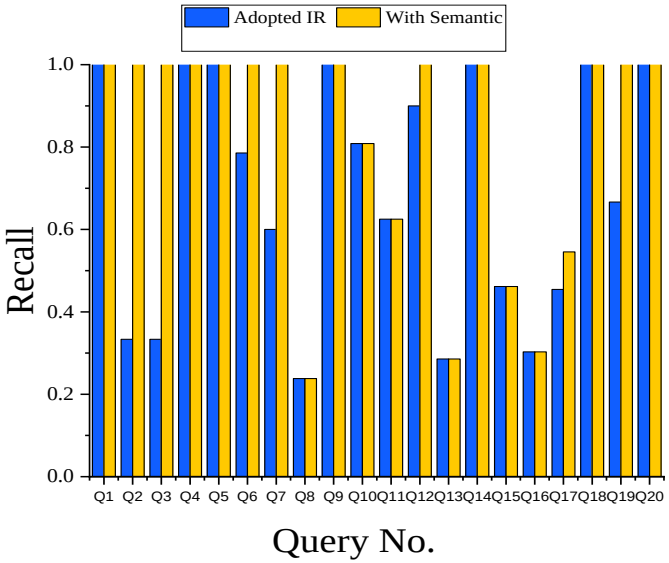
Query Expansion using Semantics of Keywords (cont.)

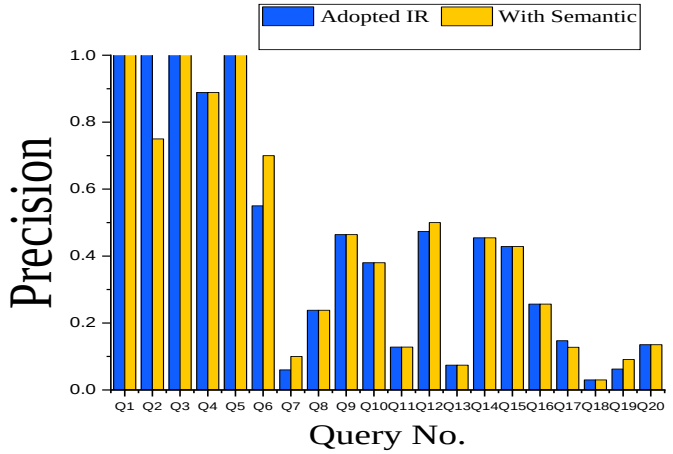
Expansion of User Query using Synonyms/Semantics

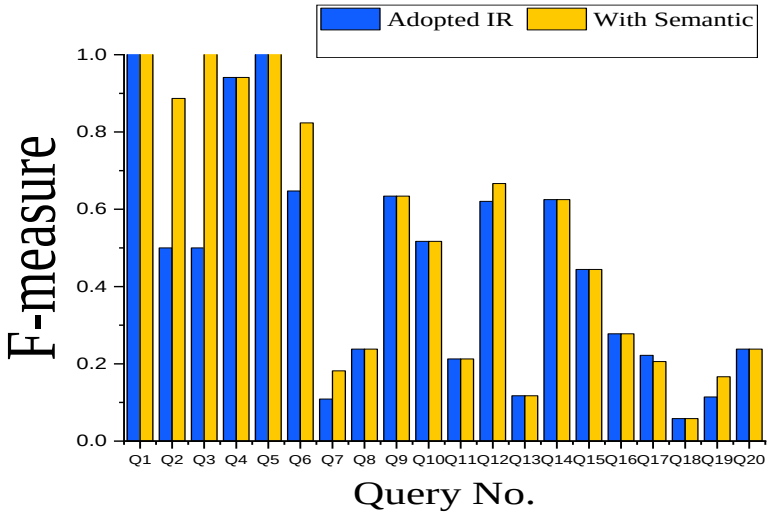
Initial Query	Expanded Query
زراعة العنب	زراعة، فلاحة ، العنب ، الكرم
تسميد الكرنب	تسميد، تخصيب ، الكرنب ، الملفوف
إنتاج البلح	إنتاج، البلح، التمر

❑ Implementation and Experimental Results (cont.)

➢ Adopted IR and Semantic Keywords







Improve 27%

❑ Third Approaches (cont.)

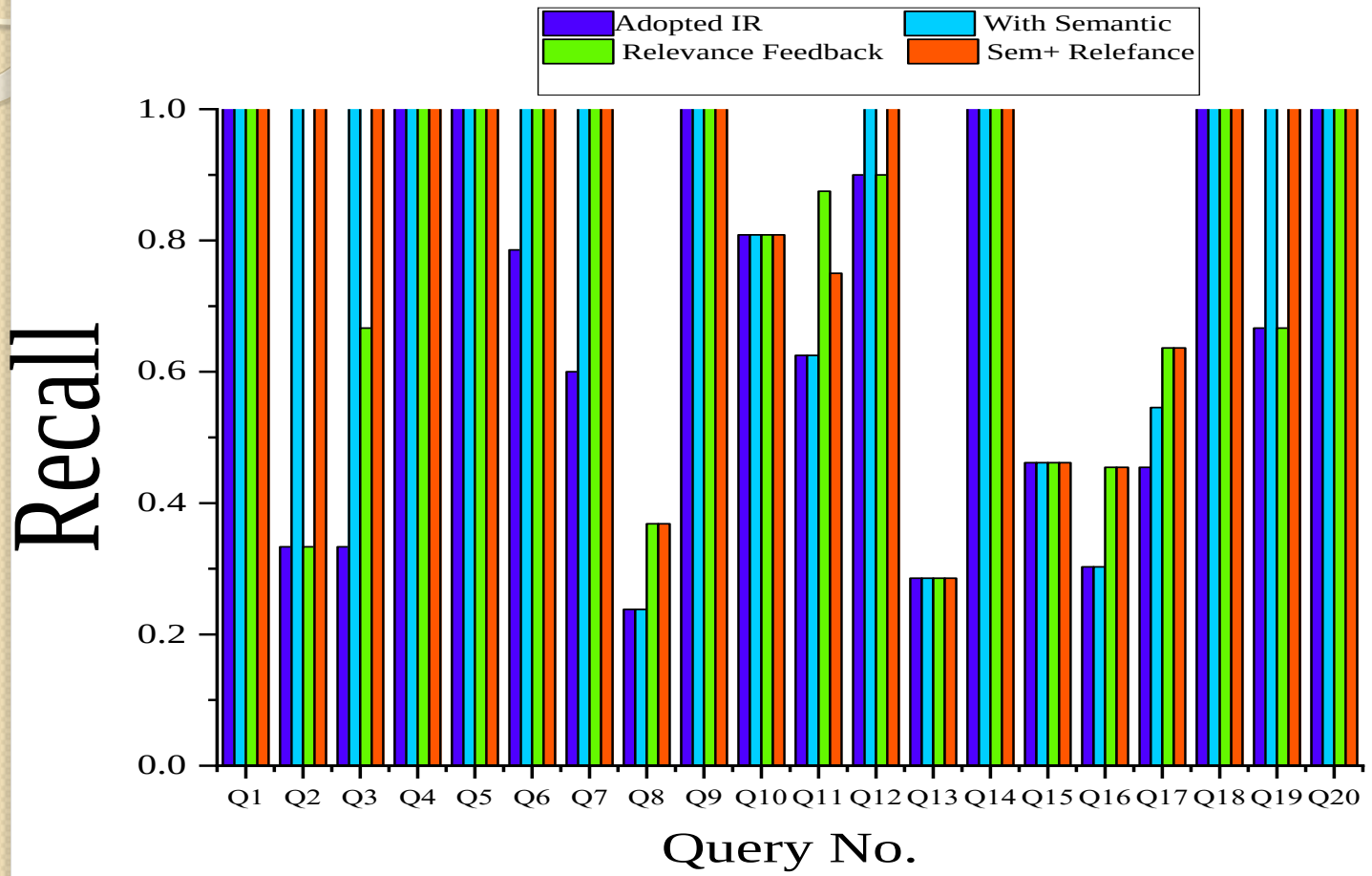
➢ Semantics of Words and Relevance Feedback Approach

- In this approach the query is modified by using semantic word and relevance feedback.

$$q_{exp} = q_{user} + \sum_{di \in s_1} d_i + S(q_r)$$

❑ Implementation and Experimental Results (cont.)

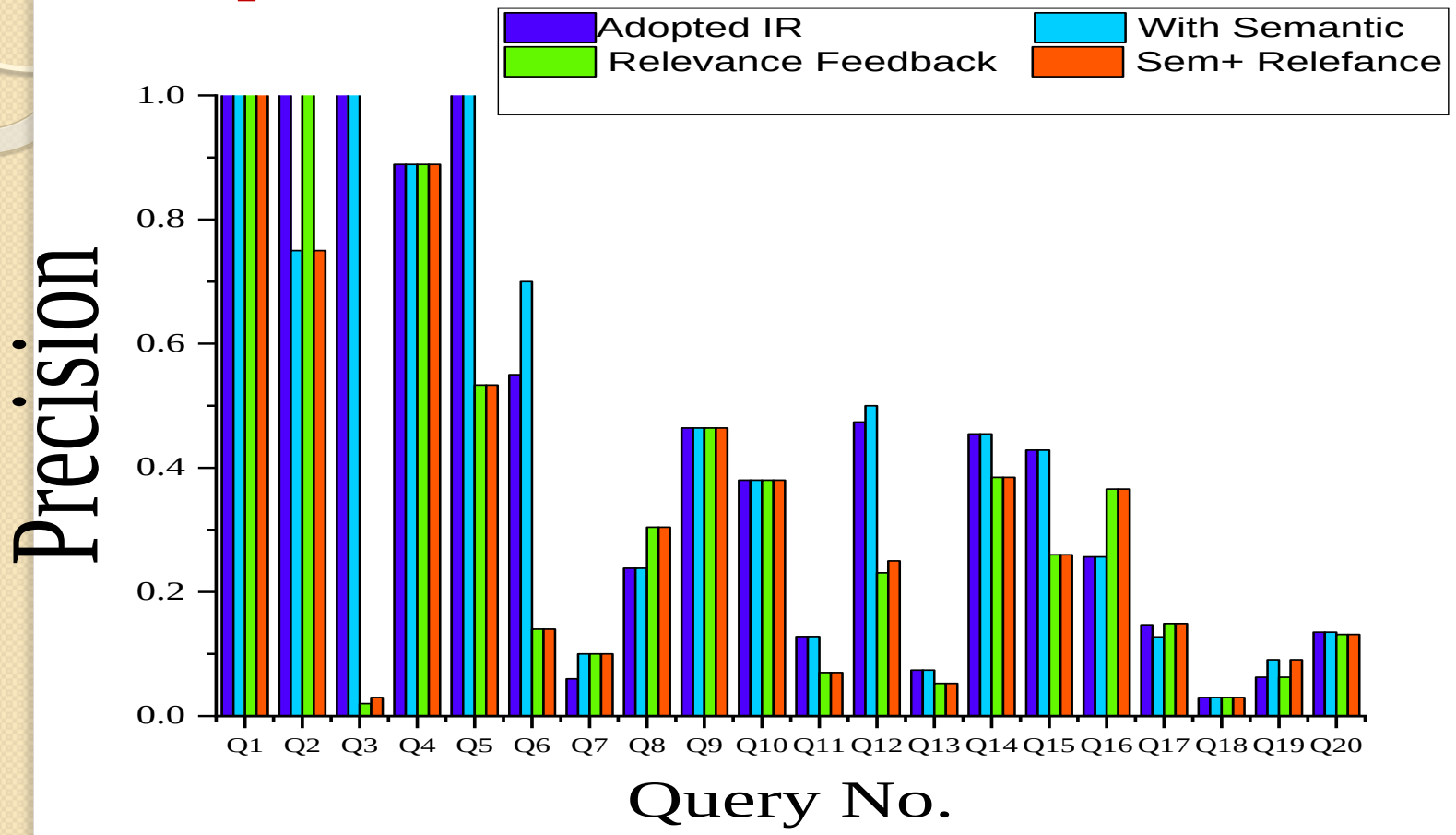
➢ Adopted IR, Relevance, Semantic and Both



Improve 15-35%

❑ Implementation and Experimental Results (cont.)

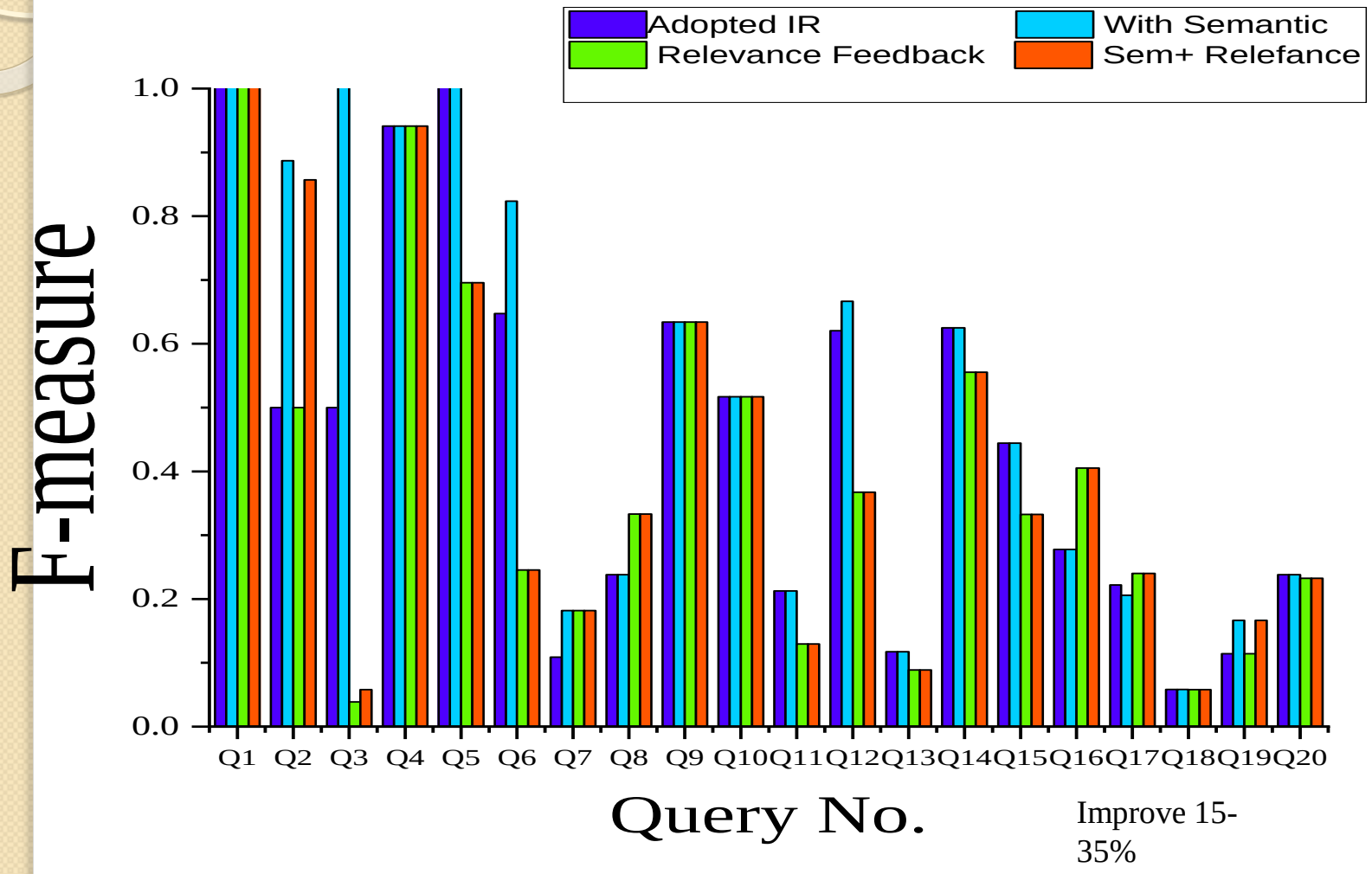
➢ Adopted IR, Relevance, Semantic and Both (cont.)



Improve 15-35%

❑ Implementation and Experimental Results (cont.)

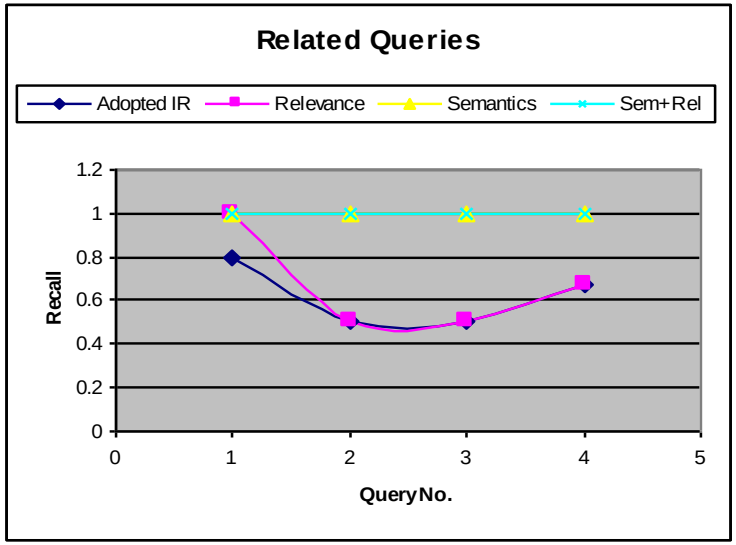
➢ Adopted IR, Relevance, Semantic and Both (cont.)

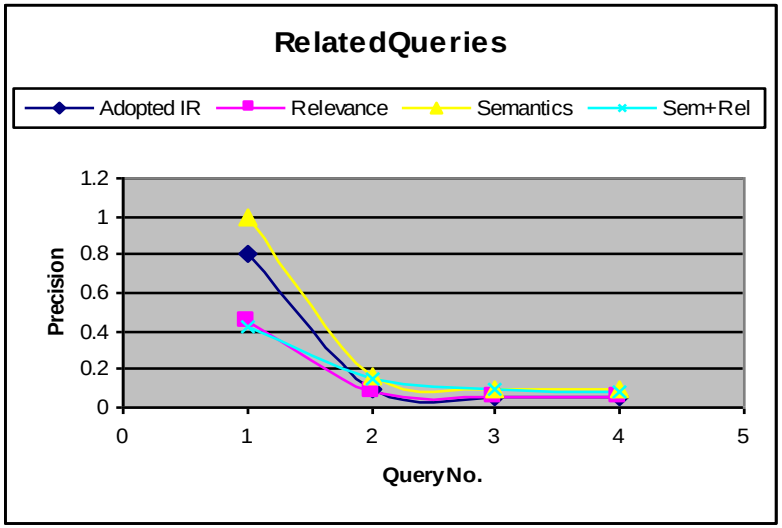


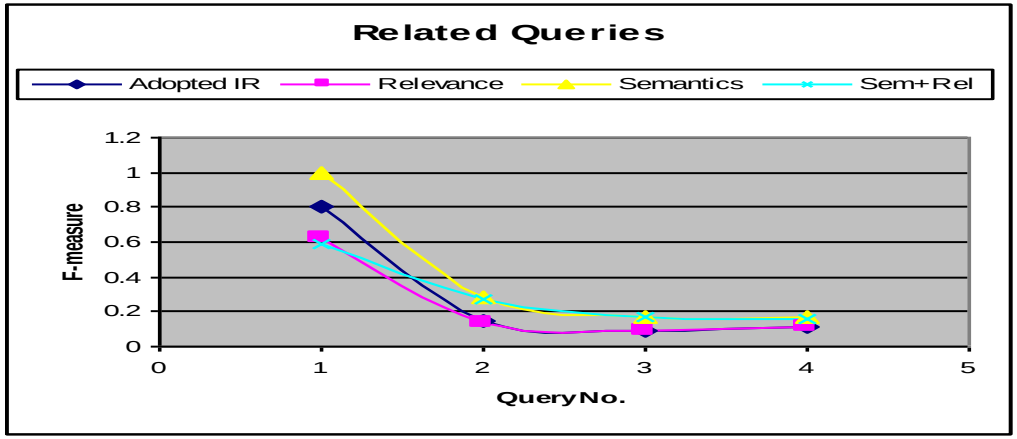
Improve 15-35%

❑ Implementation and Experimental Results (cont.)

➢ Adopted IR, Relevance, Semantic and Both (cont.)





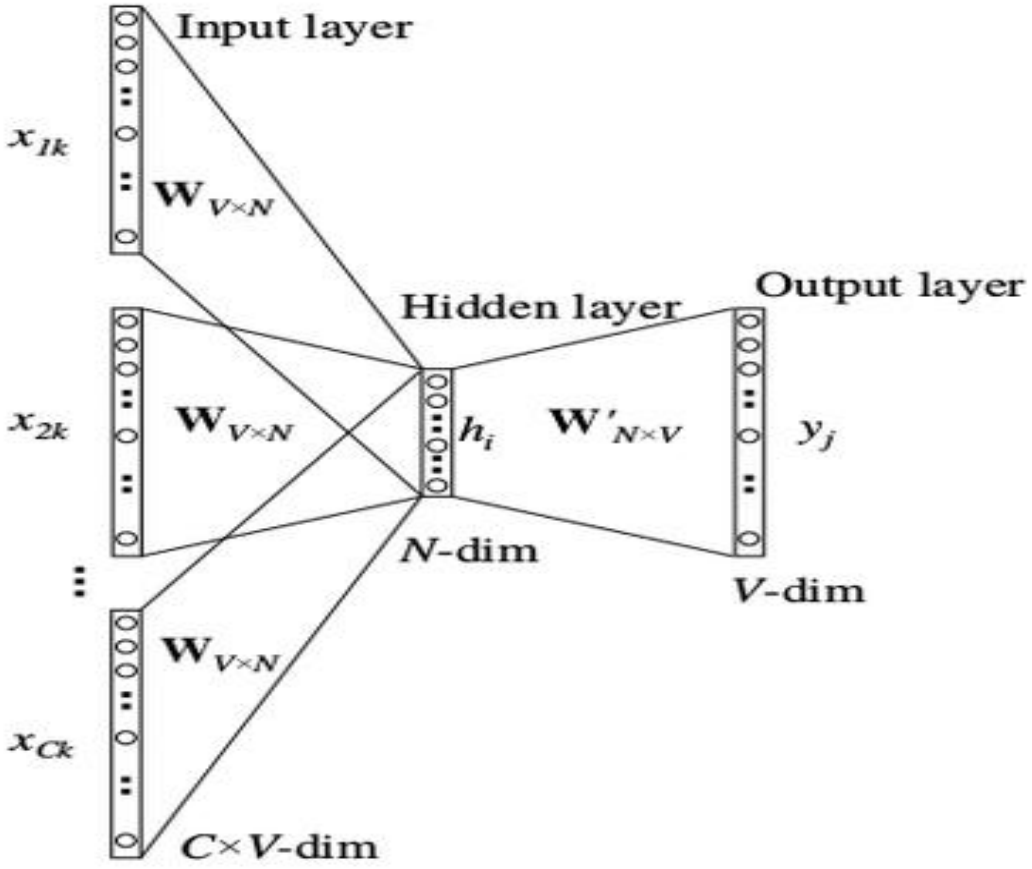


❑ Query Expansion Using the Word Embedding

- **Word embedding** is a form of word representation that learned representations of text in an n-dimensional space.
- Meaning that two similar words are represented by almost similar vectors that are very closely placed in a vector space.
- **Word2Vec** is a statistical method for efficiently learning a standalone word embedding from a text corpus.
- It was developed by Tomas Mikolov, et al. at Google as a response to make the neural-network-based training of the embedding more efficient.
- There are two different learning models
- Continuous Bag-of-Words (CBOW) model.
- Skip-Gram Model.

Continuous Bag-of-Words (CBOW) model

- The CBOW model learns the embedding by predicting the current word based on its context.

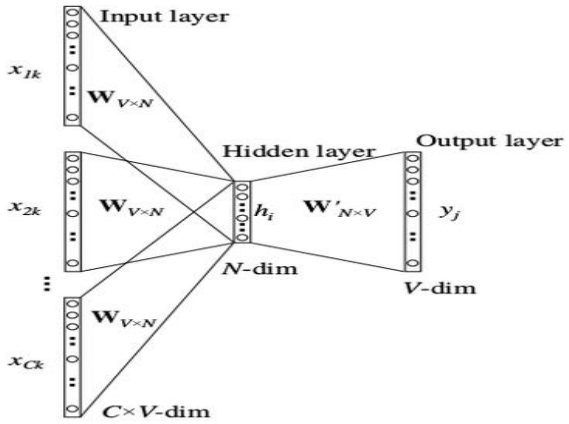


CBOW model (cont.)

- $x_{1k}\ldots.x_{ck}$ are the input context words
- V is the vocabulary size
- N is the size of the hidden layer
- W_{VXN} is The weights between the input and the hidden layer
- The hidden layer of this CBOW model is calculated by this equation

➢ $h = \frac{1}{C} W^T (x_1 + x_2 + \ldots.. + x_c)$ [3] (5)

$= \frac{1}{C}(v_{w1} + v_{w2} + \ldots.. + v_{wc})^T$

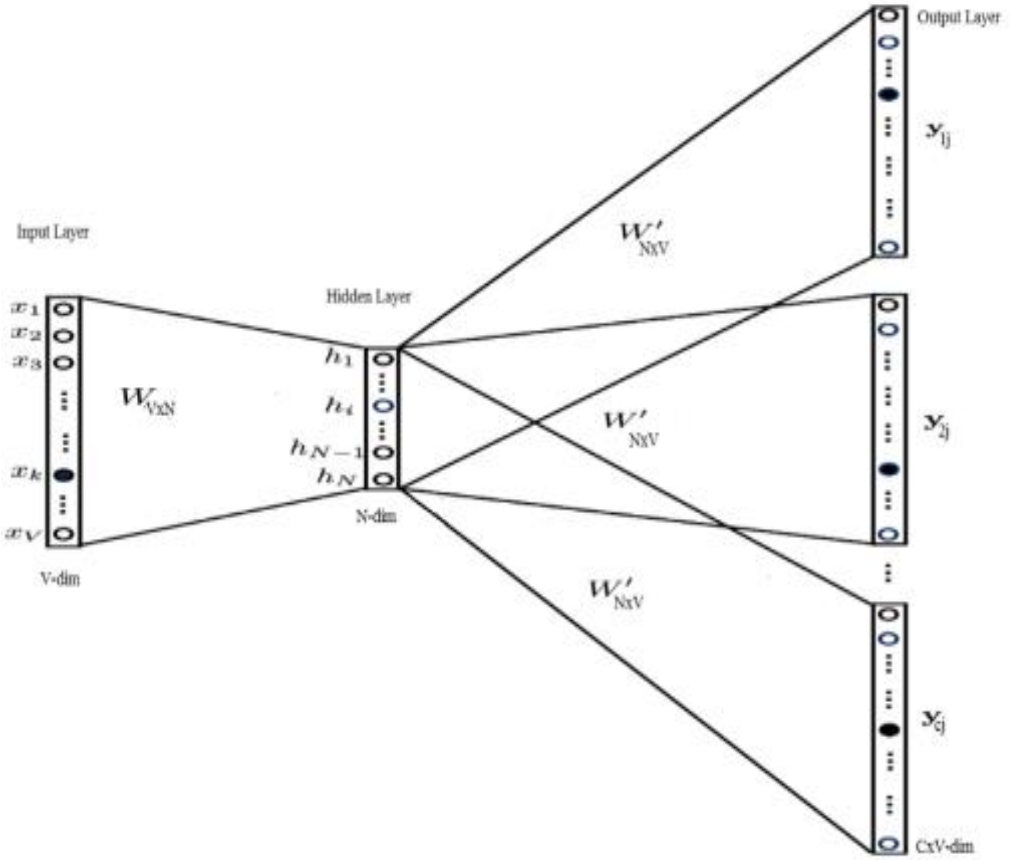


Where

- C is the number of words in the context,
- $(x_1, \ldots., x_c)$ are the inputs to the network as a one-hot encoded vector
- $w_1, \ldots, w_c$ are the words in the context
- v_w is the input vector of a word w.

❑ Skip-Gram model

- The skip-gram model learns by predicting the surrounding words given a current word.



❑ Skip-Gram model (cont.)

- Each output is computed using the same hidden-output matrix:

$$P(w_{c,j} = w_{o,c}|w_I) = y_{c,j} = \frac{\exp(u_{c,j})}{\sum_{J'=1}^{V} \exp(u_{J'})}$$

Where

- $w_{c,j}$ is the j^{th} word on the c^{th} panel of the output layer
- $w_{o,c}$ is the actual c^{th} word in the output context words
- w_I is the only input word
- $y_{c,j}$ is the output of the j^{th} unit on the c^{th} panel of the output layer
- $u_{c,j}$ is the net input of the j^{th} unit on the c^{th} panel of the output layer.

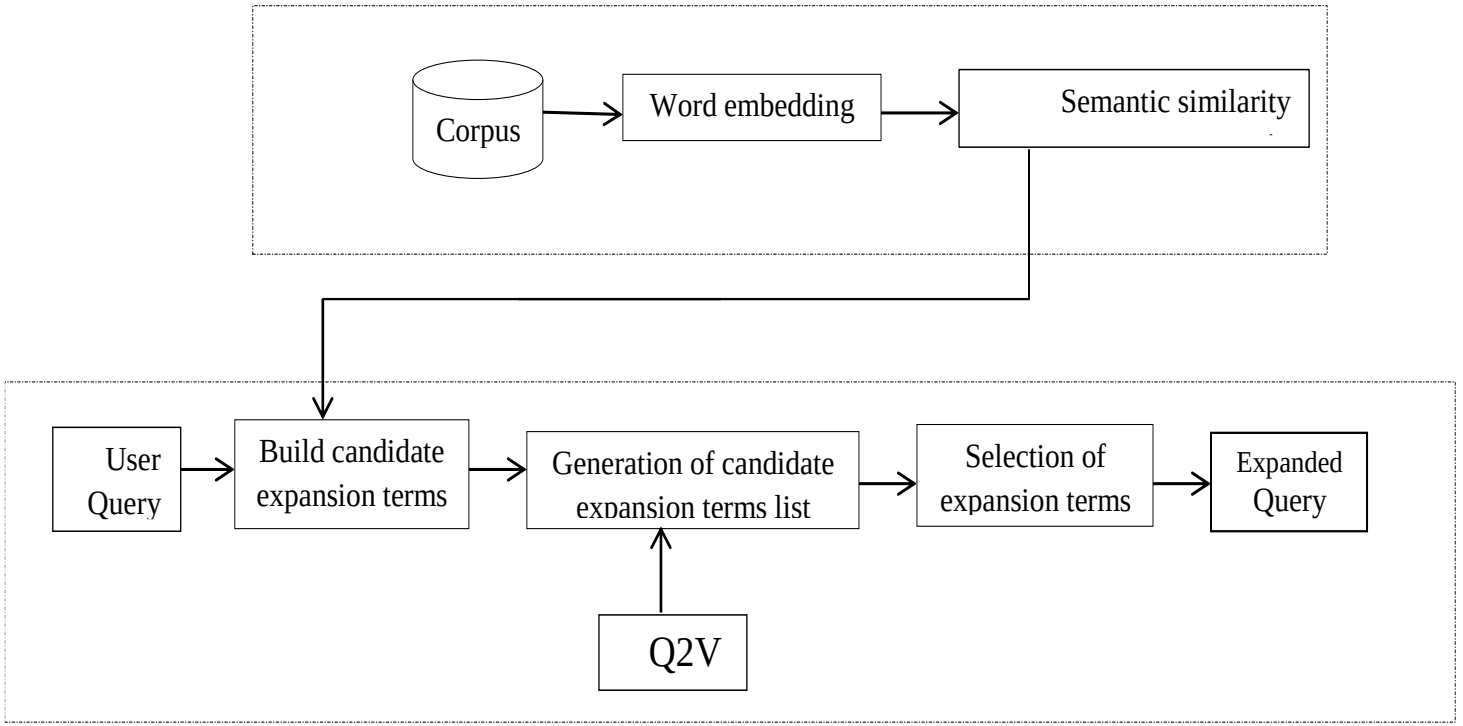
- Because the output layer panels share the same weights, thus the net input can be written as

$$u_{c,j} = u_i = V'^{T}_{w_j}.h \qquad \text{for c=1,2,….,C}$$

Where

- V'_{wj} the output vector of the j^{th} word in the vocabulary

❑ Analysis of a Word Embedding Approach



❑ Analysis of a Word Embedding Approach (cont.)

- Assume that the user query is represented as a set Q which contains all the query terms; i.e Q = {q_1 , q_2 ,…., q_n }
- Assume that W is the set of all vocabulary terms related to the query terms.
- Assume that V is the subset of vocabulary terms that doesn't contain the query terms. i.e V=W-Q and V={t_1, t_2,…t_m}.
- Assume matrix E contains the similarity values between all query terms q_1, q_2, … q_n and the candidate terms t_1, t_2 … t_m. the E matrix is represented as:

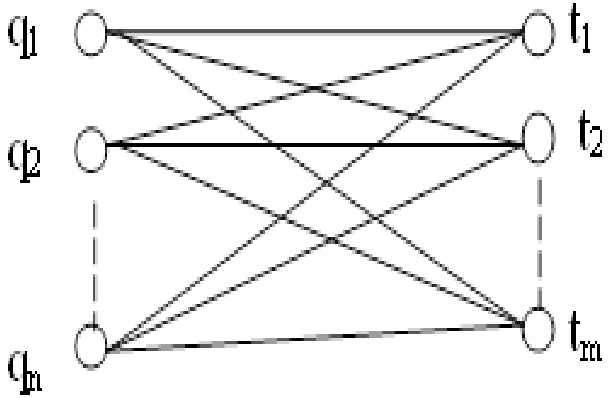
E= {cos (q_1,t_1), cos (q_1,t_2), ...cos (q_2,t_1), cos (q_2,t_2), … cos (q_n,t_m)}

- The query expansion process can be represented by Q2V approach.

❑ Analysis of a Word Embedding Approach (cont.)

▪ The Query Guided Q2V Approach

- This approach aims to select the top closest prospects for expansion.
- Assume that f (q, Q) is the frequency of a term q in the query Q.
- Every query term q ∈Q gives a vote to all candidate terms t ∈ V constructing that is called weighted cosine similarity w.



❑ Analysis of a Word Embedding Approach (cont.)

▪ The Query Guided Q2V Approach (cont.)

- Each query term q constructs a list; $Rank_q$; of ordered candidate-vote pairs as shown in equations

$$w(q, t) = f(q, Q) \cos(q, t) \qquad (7)$$

$$\forall q \in Q, Rank_q = \{(t, w(q, t)) : q \in Q, \forall t \in V$$

- Select the best or top N candidate terms ranked by the semantic cosine similarity. N may be 1, 2, 3 or more. This means that all query terms in the user query select at least one term of the candidate terms in V.

❑ Implementation and Experimental Results

➢ Document Collection Dataset

- The dataset is taken from the CNN English Website
- It contains about 92000 news documents with a total size 6 GB.

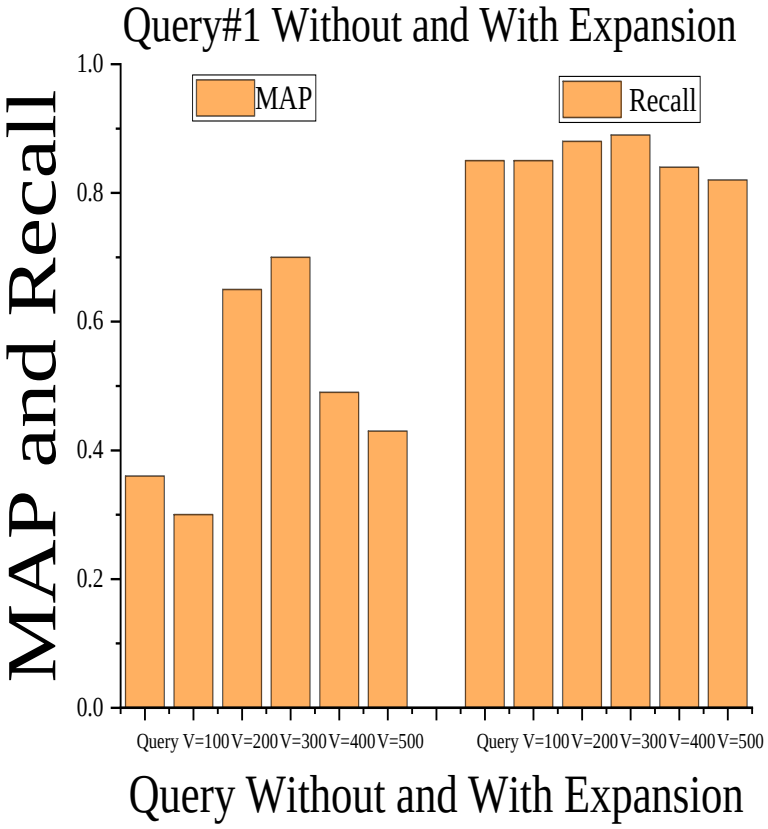
➢ Performance Metrics

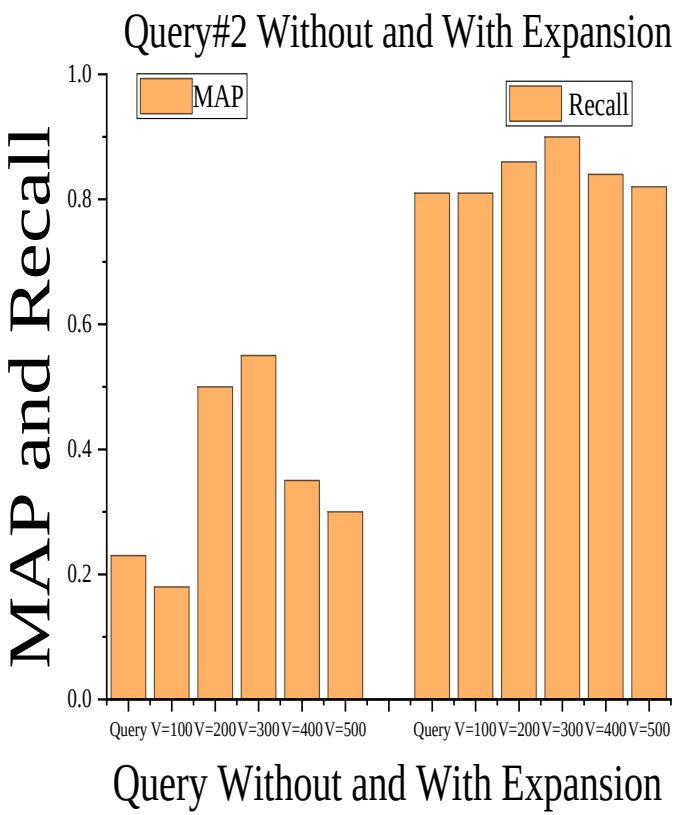
- **Recall**
- **MAP**

- MAP The mean average precision
- MAP=$\sum_{q=1}^{Q} \frac{AveP(q)}{Q}$

❑ Implementation and Experimental Results (cont.)

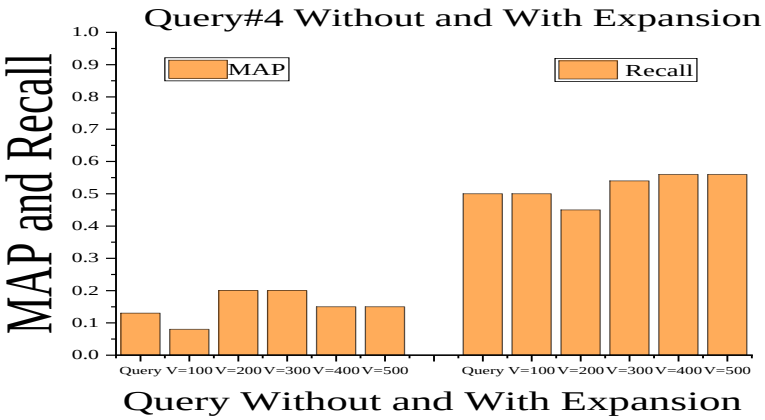
➢ Effect of Vector Length

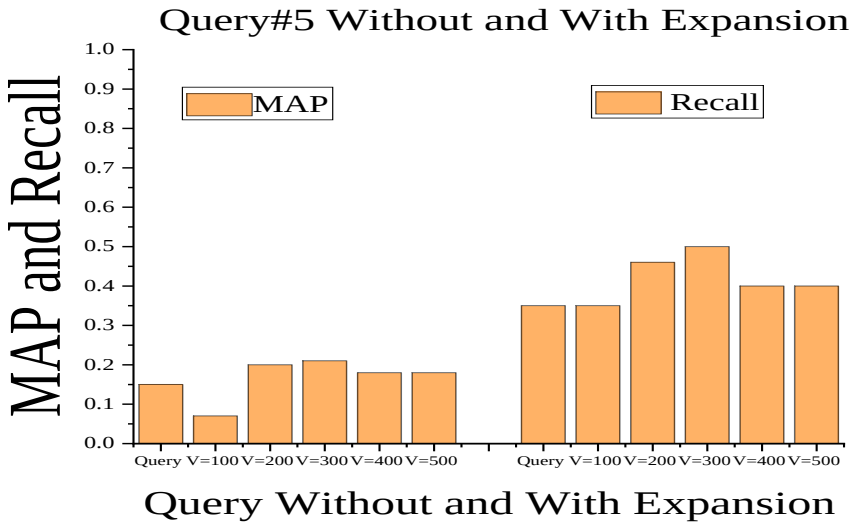




❑ Implementation and Experimental Results (cont.)

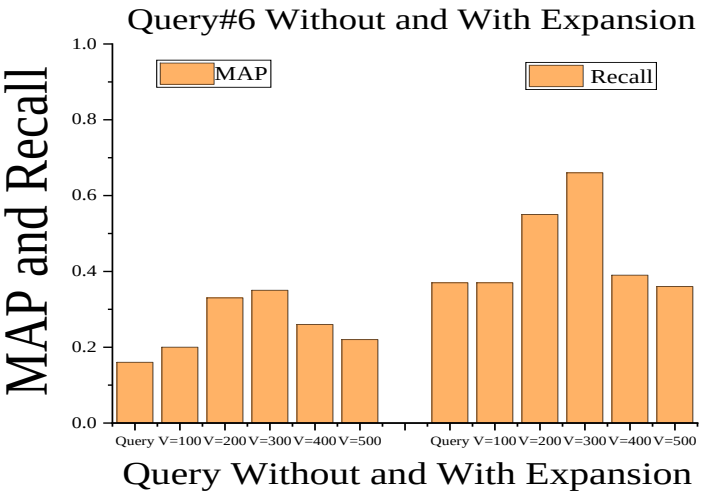
➢ Effect of Vector Length (cont.)

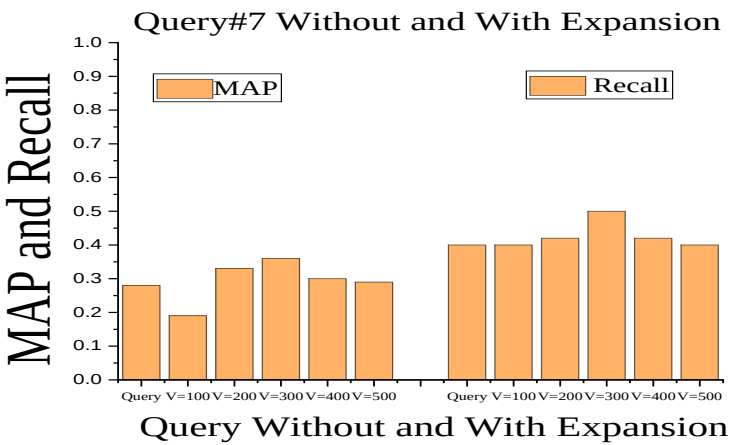


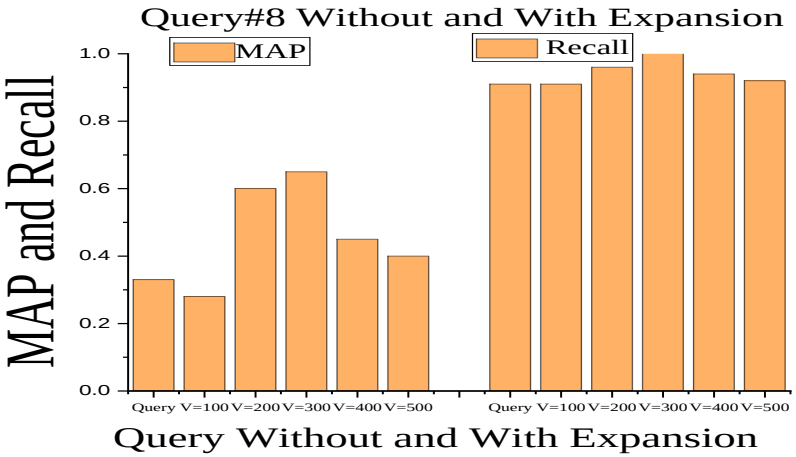


❑ Implementation and Experimental Results (cont.)

➢ Effect of Vector Length (cont.)

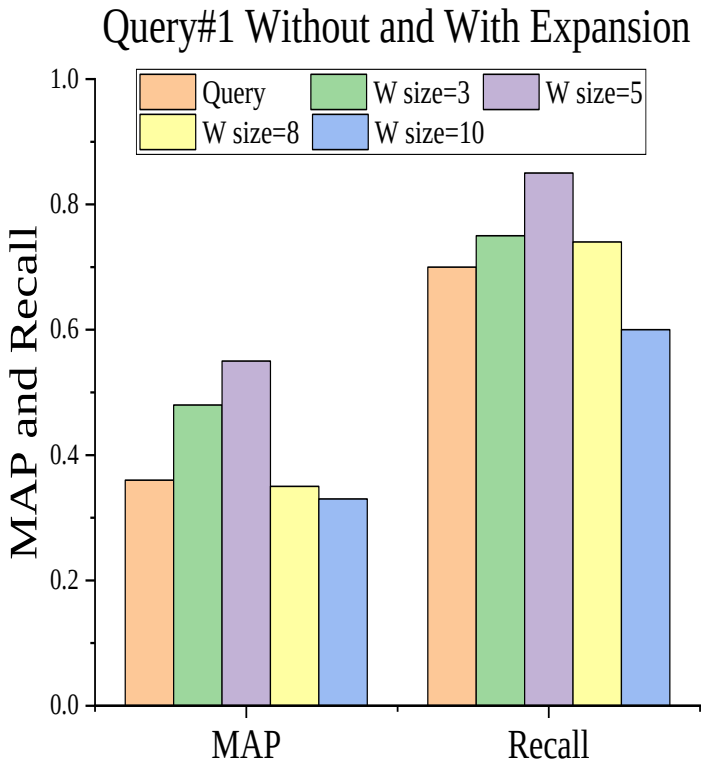


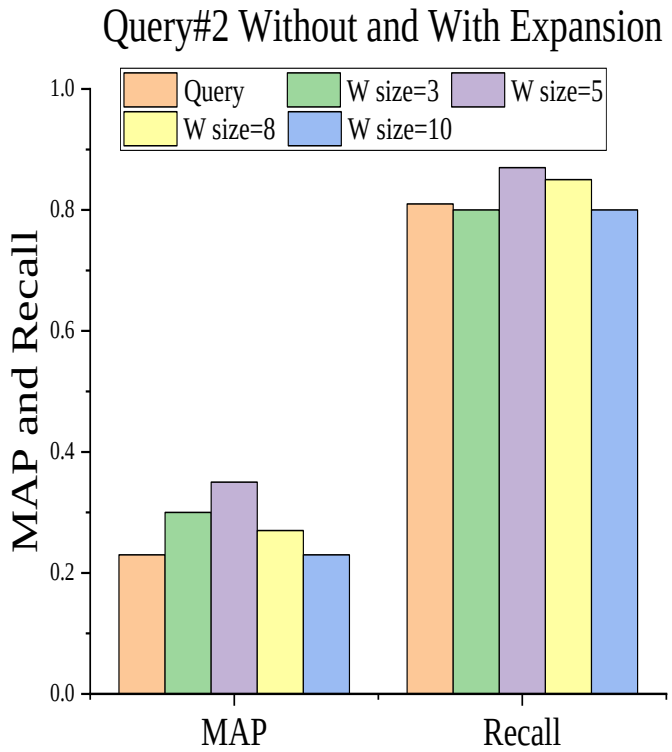




❑ Implementation and Experimental Results (cont.)

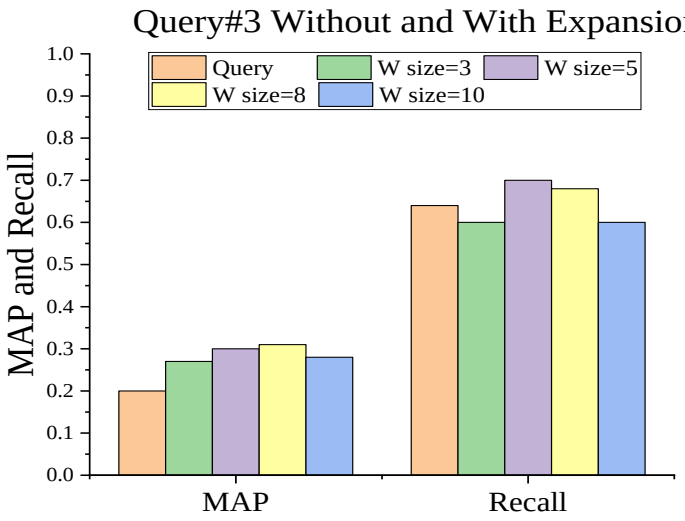
➢ Effect of Window Size

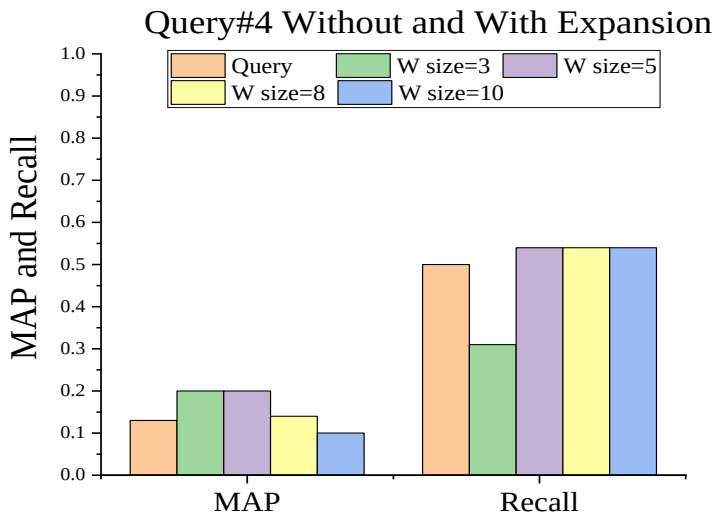


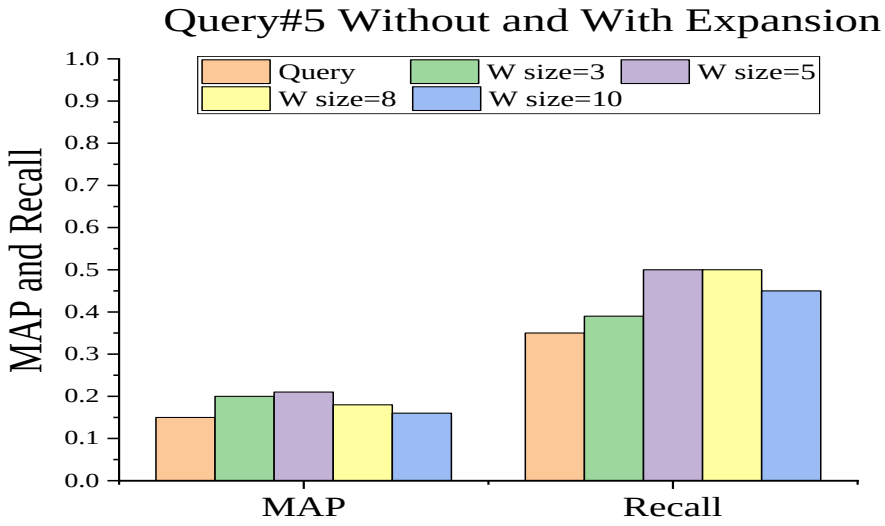


❑ Implementation and Experimental Results (cont.)

➢ Effect of Window Size (cont.)

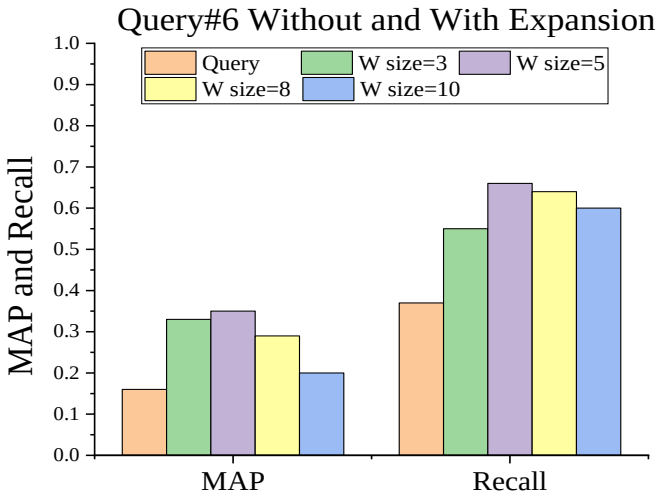


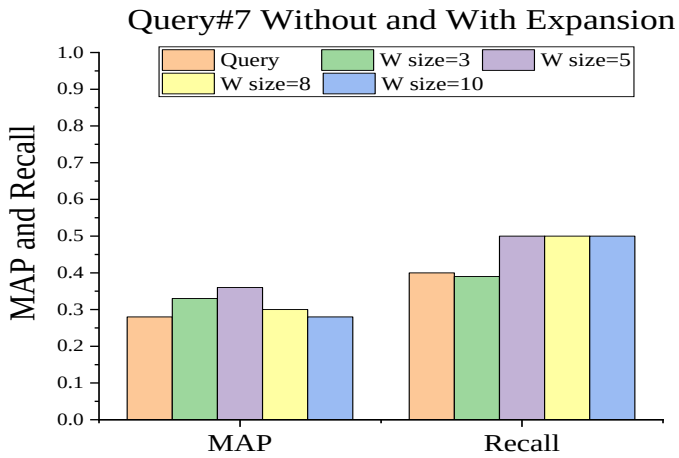


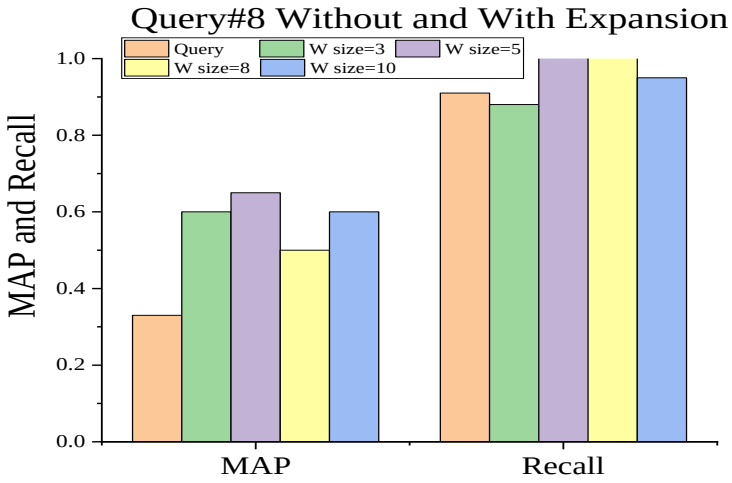


❑ Implementation and Experimental Results (cont.)

➢ Effect of Window Size (cont.)

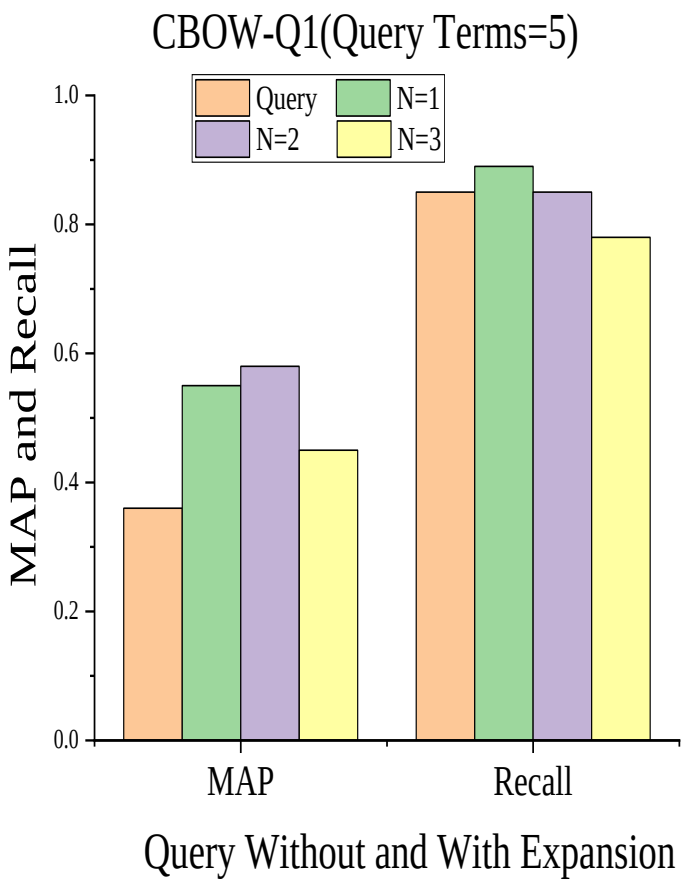


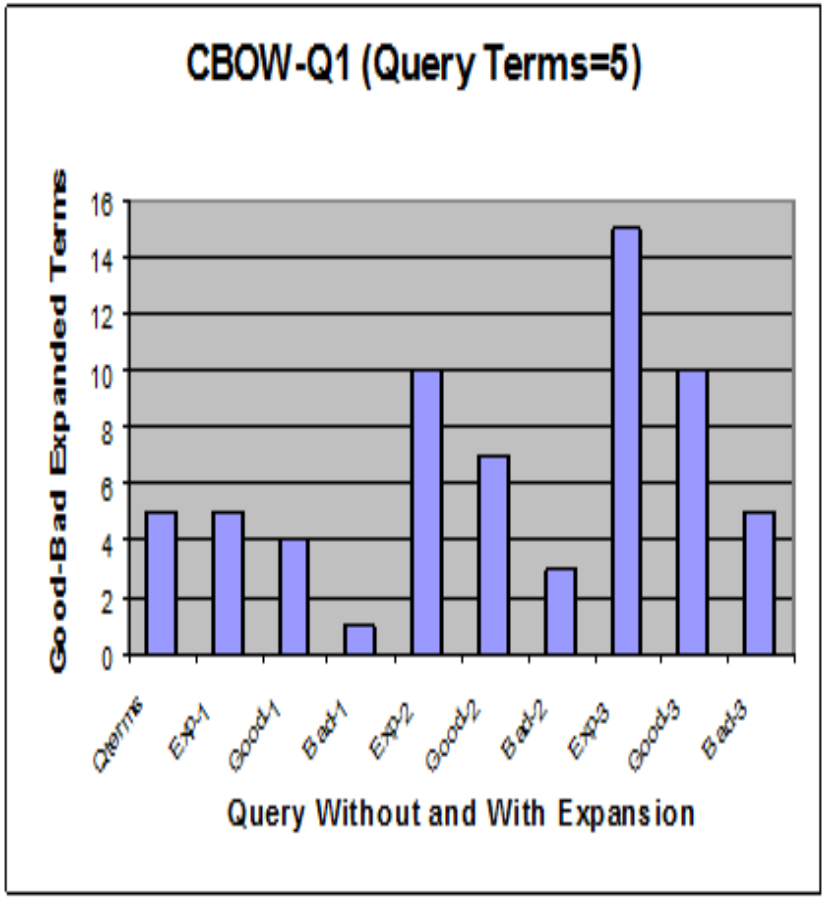




❑ Implementation and Experimental Results (cont.)

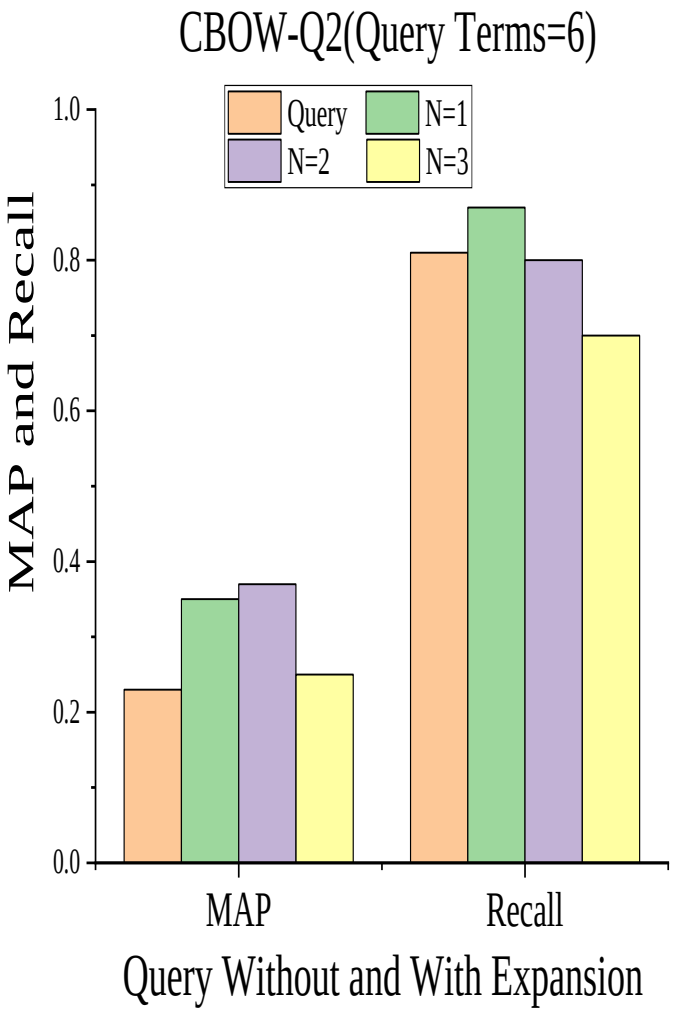
➢ Expansion Using CBOW

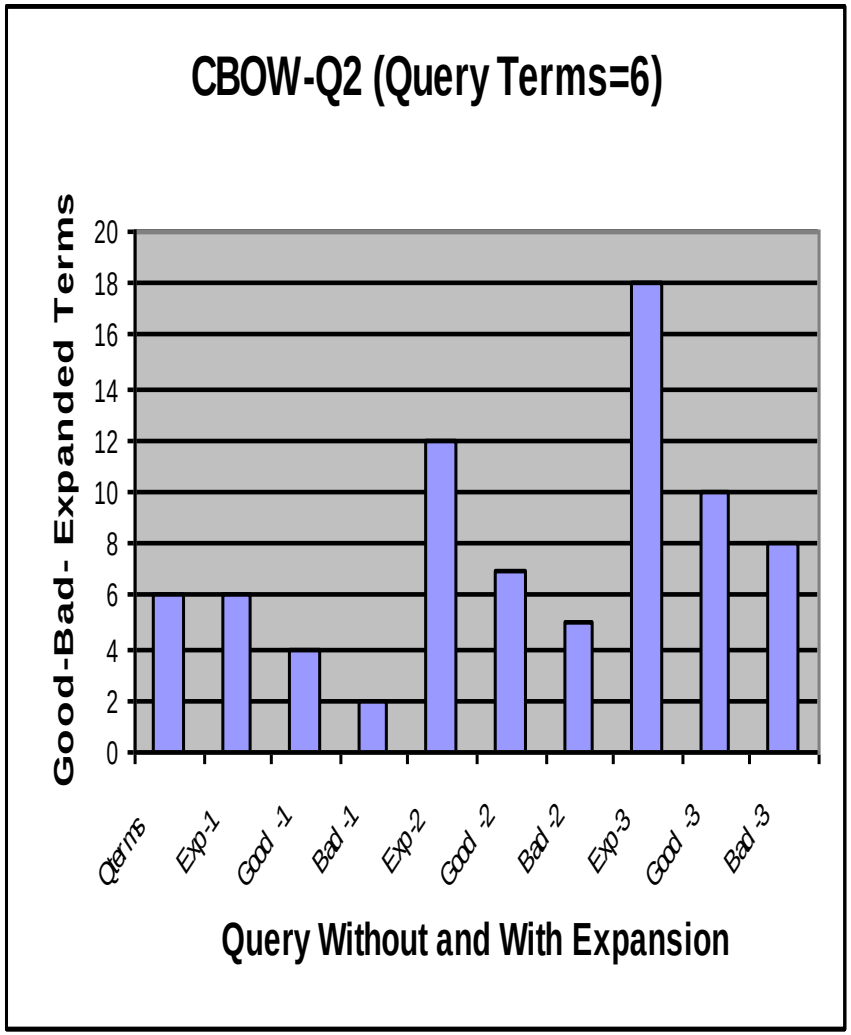




❑ Implementation and Experimental Results (cont.)

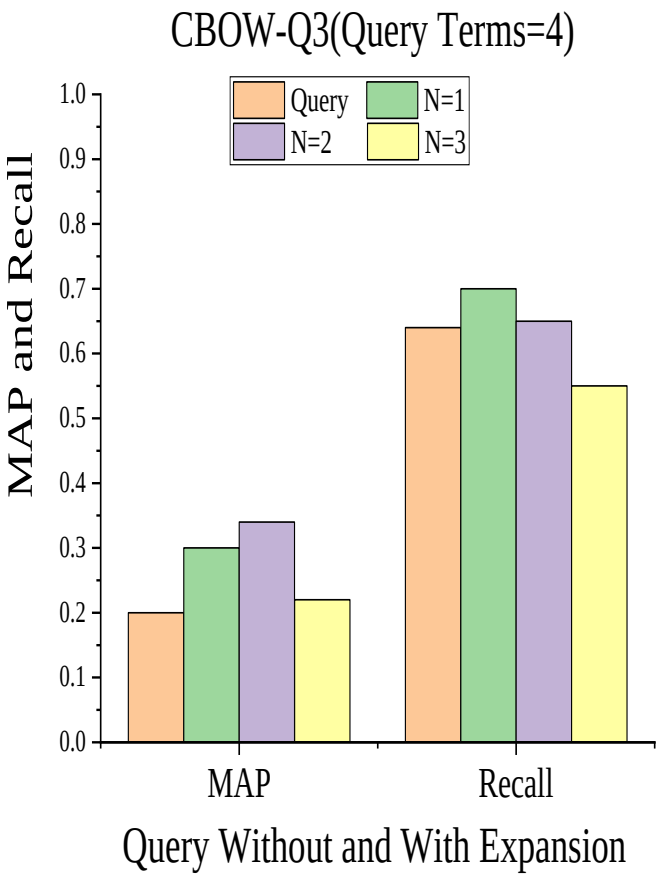
➢ Expansion Using CBOW (cont.)

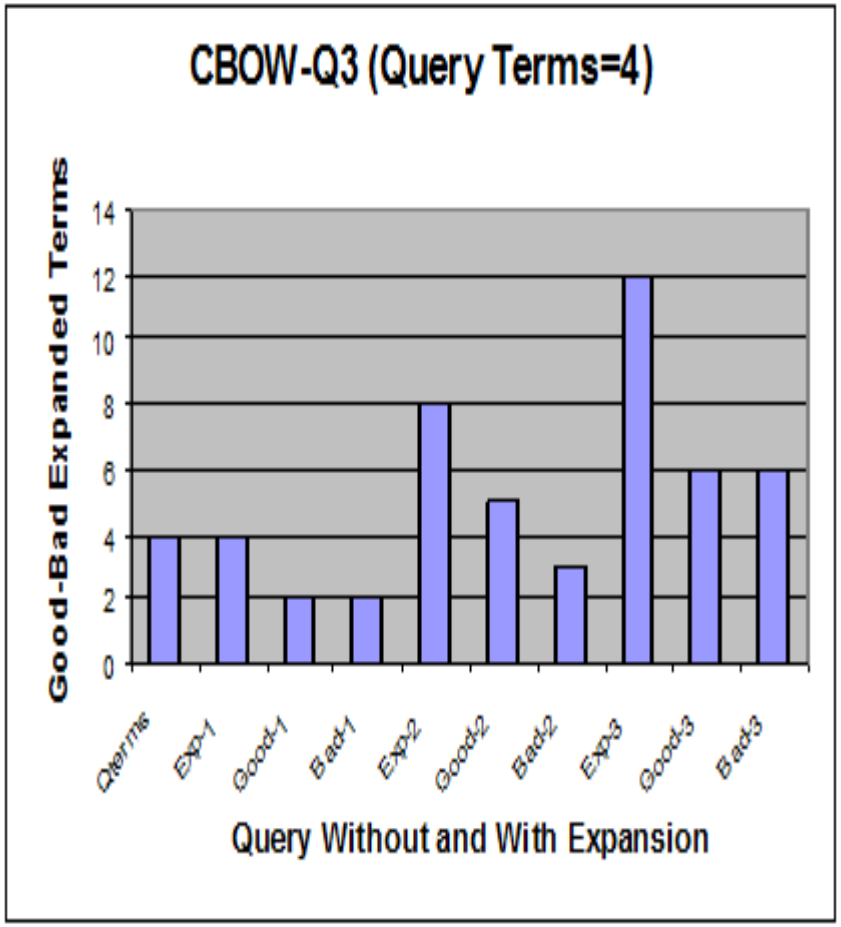




❑ Implementation and Experimental Results (cont.)

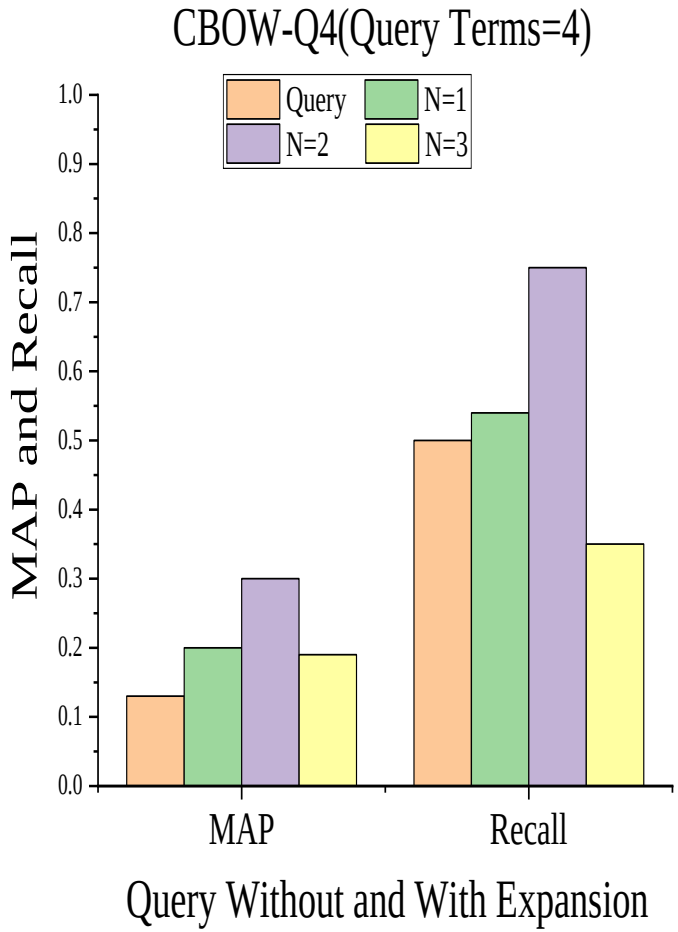
➢ Expansion Using CBOW (cont.)

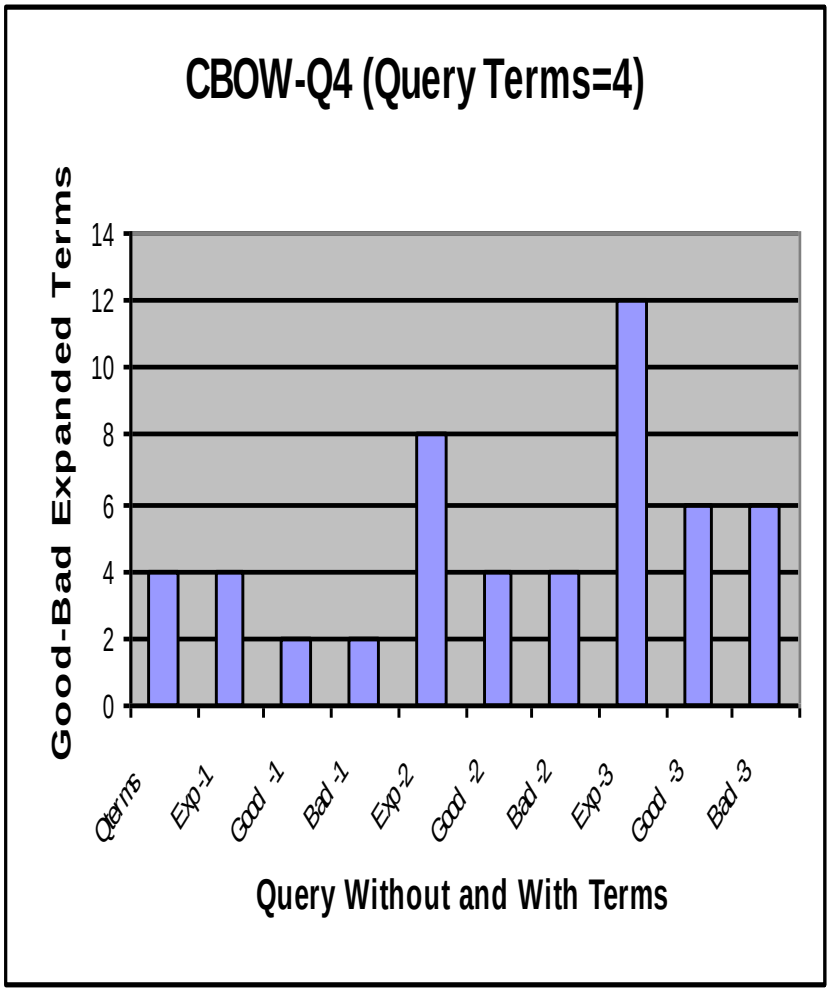




❑ Implementation and Experimental Results (cont.)

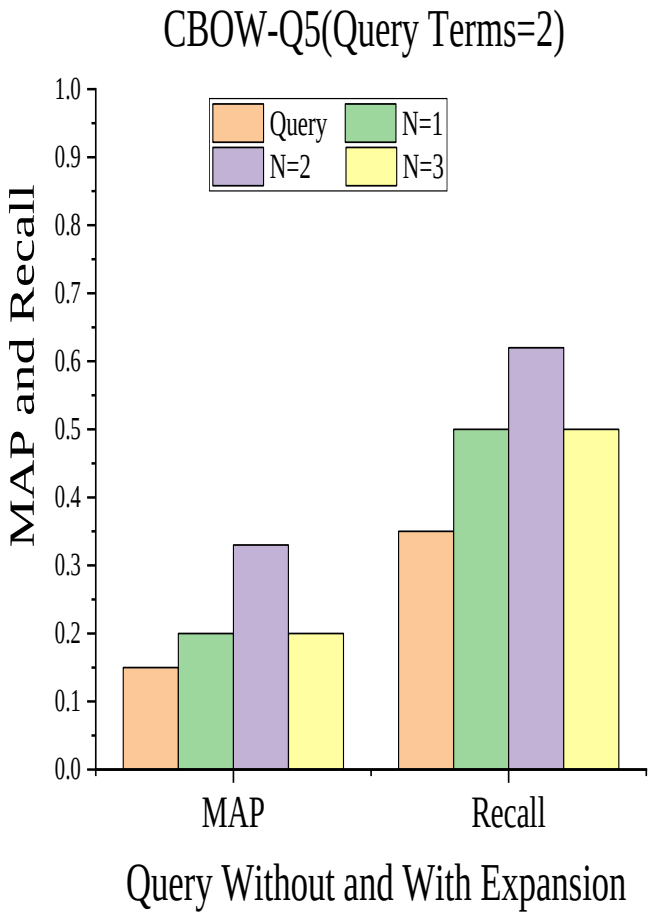
➢ Expansion Using CBOW (cont.)

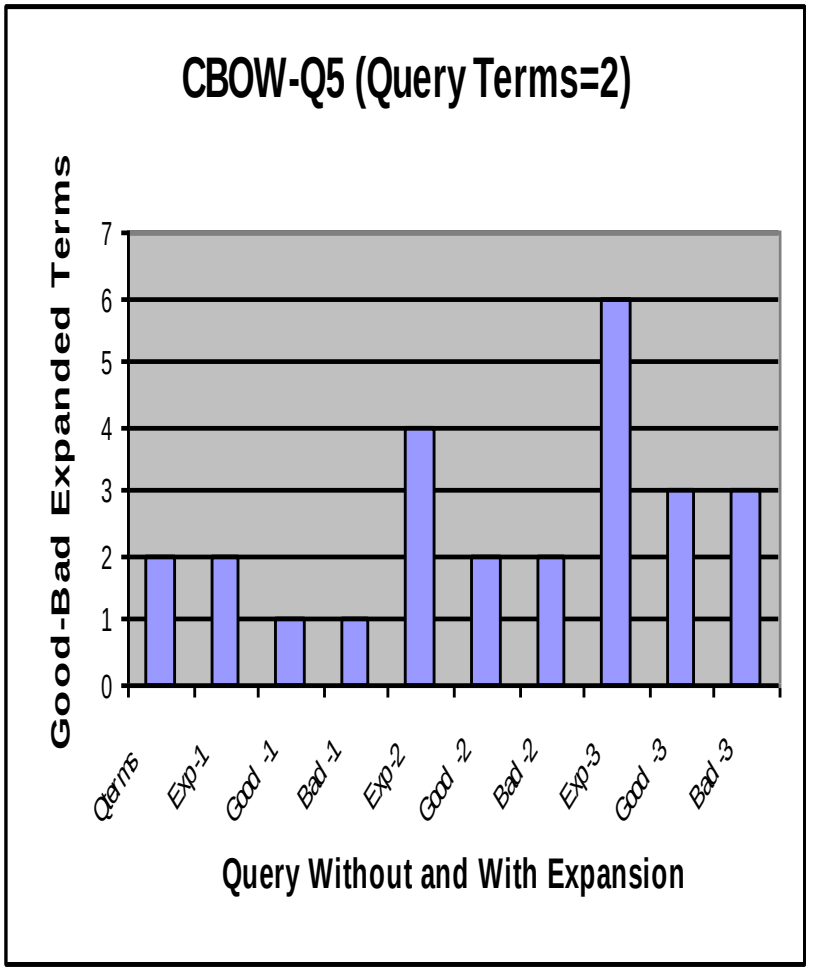




❑ Implementation and Experimental Results (cont.)

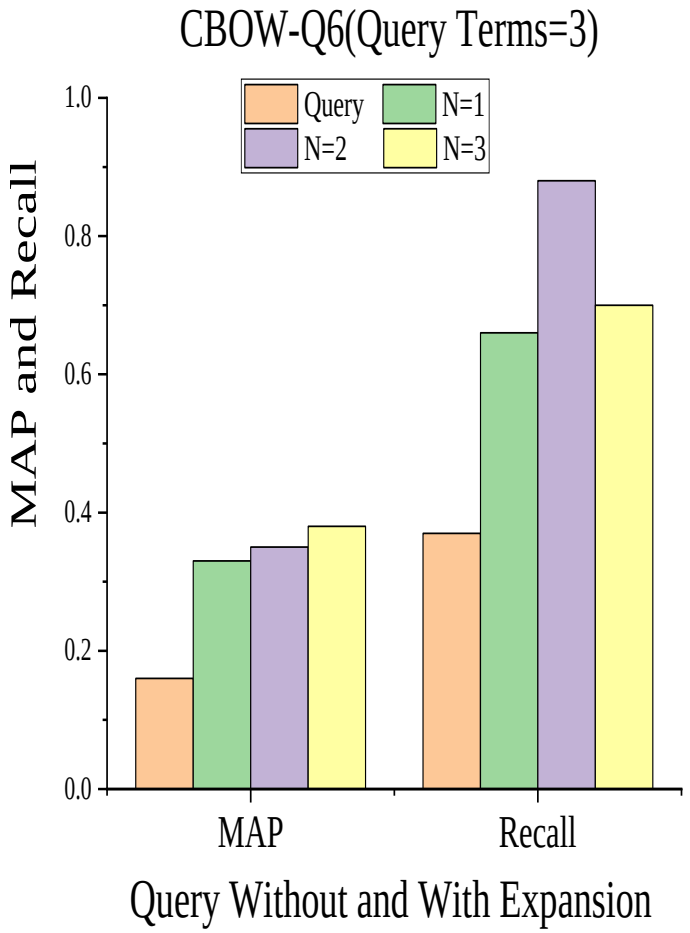
➢ Expansion Using CBOW (cont.)

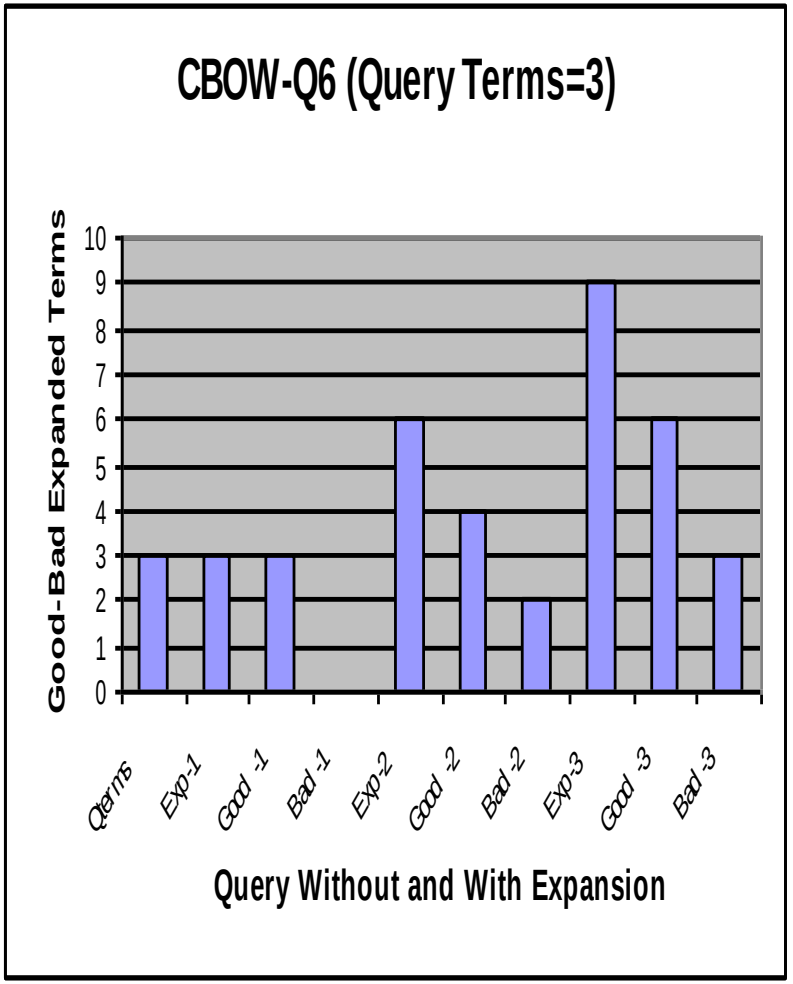




❑ Implementation and Experimental Results (cont.)

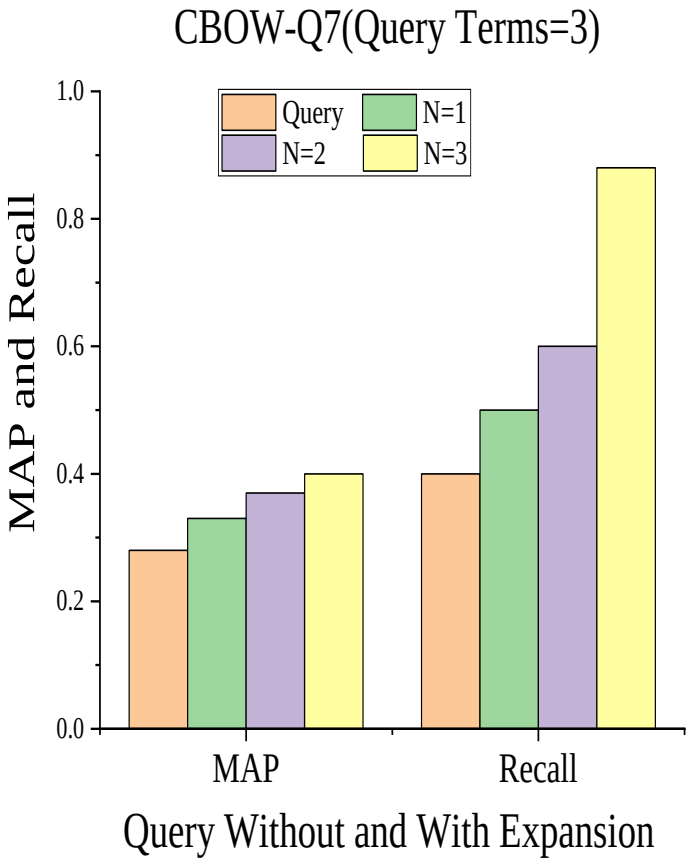
➢ Expansion Using CBOW (cont.)

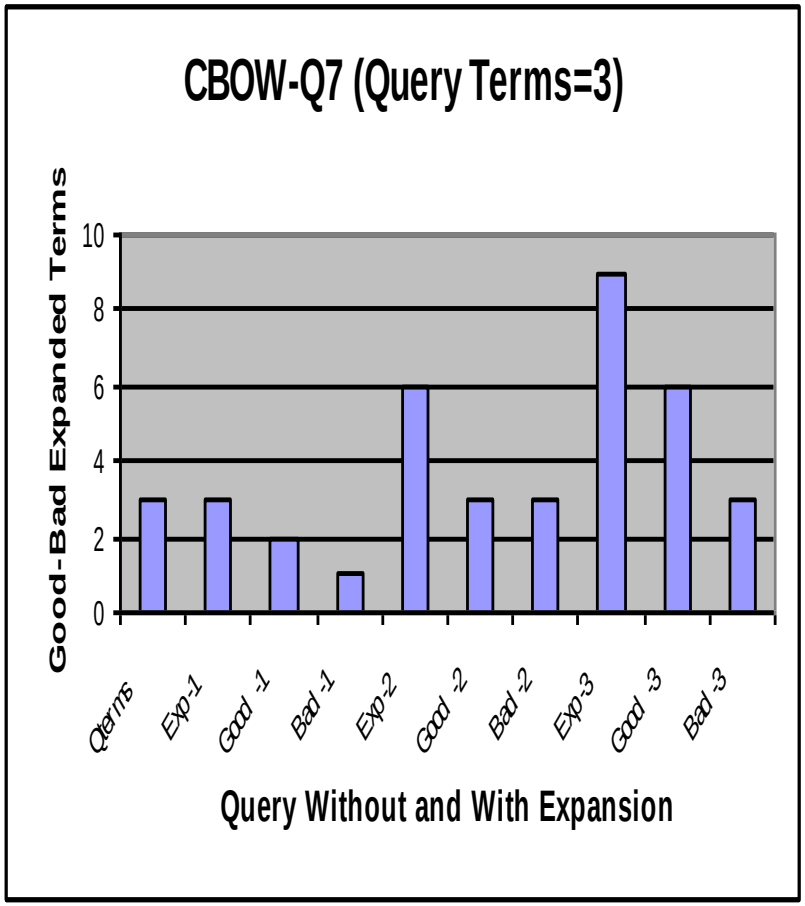




❑ Implementation and Experimental Results (cont.)

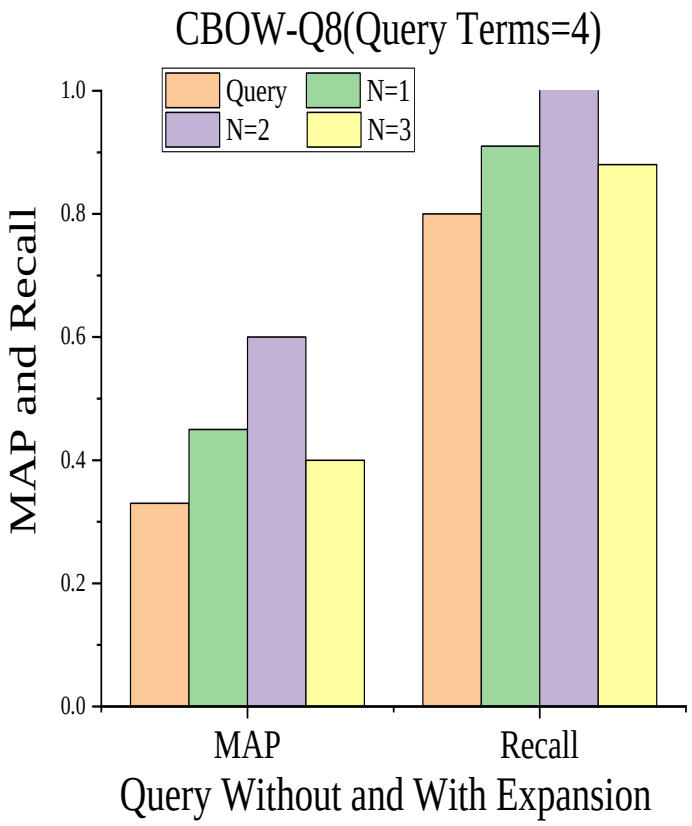
➢ Expansion Using CBOW(cont.)

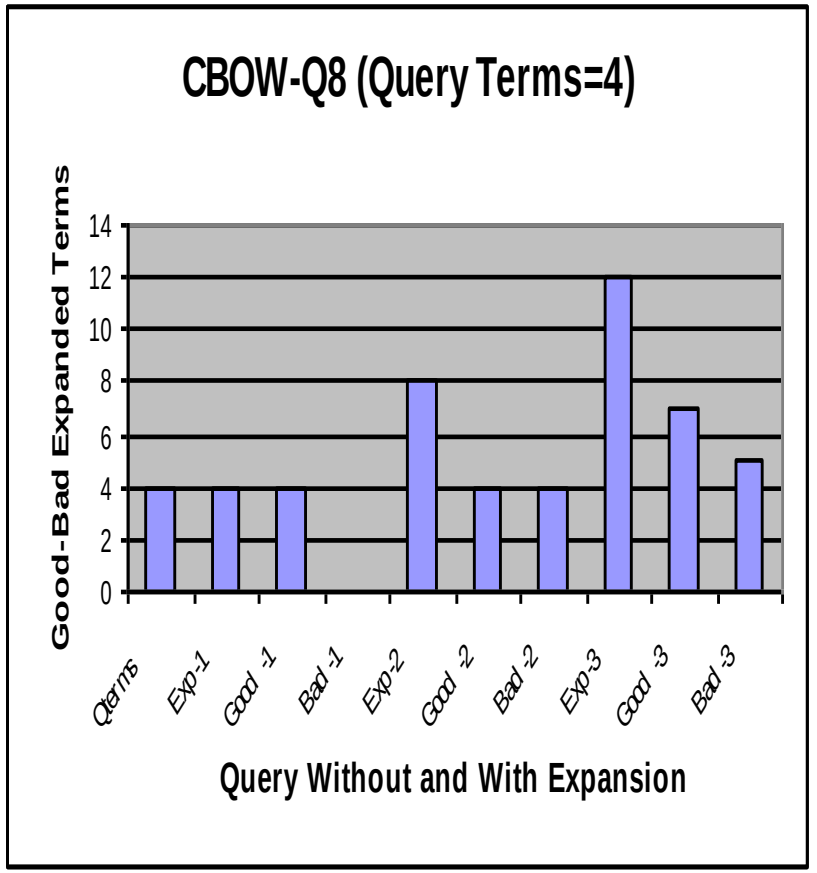




❑ Implementation and Experimental Results (cont.)

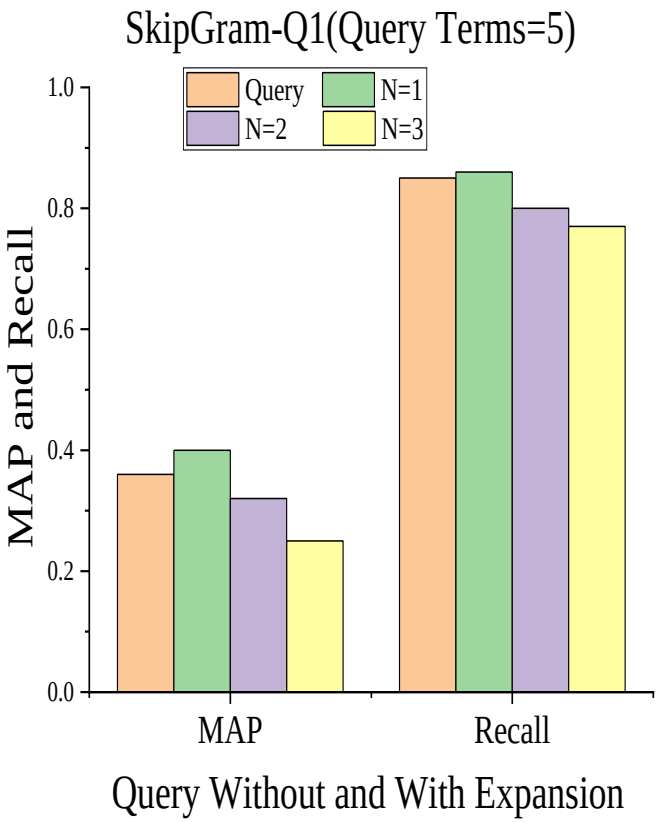
➢ Expansion Using CBOW(cont.)

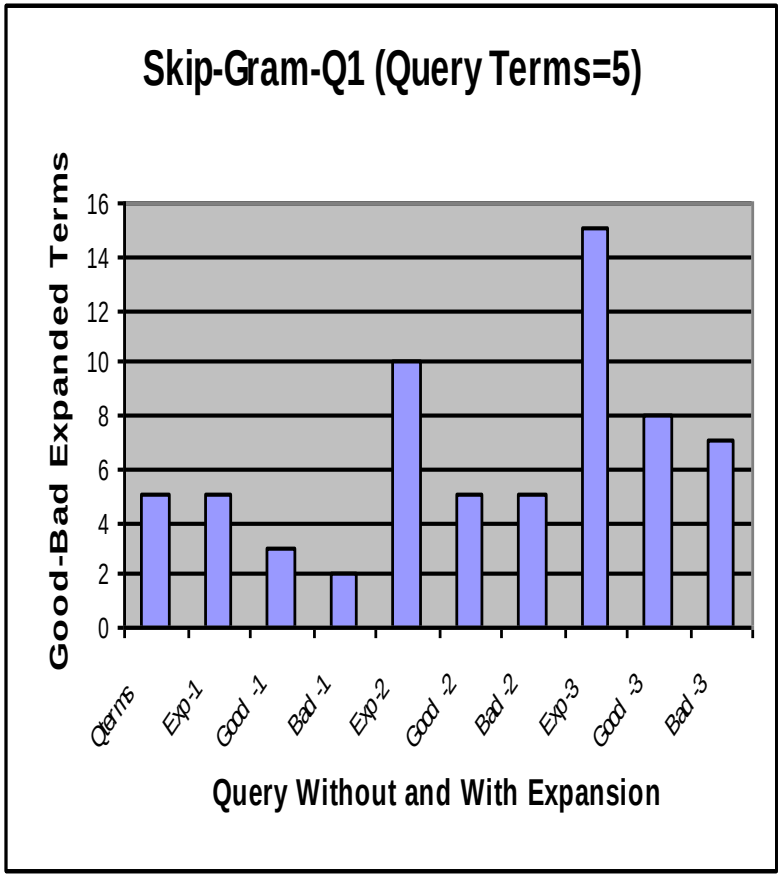




Implementation and Experimental Results (cont.)

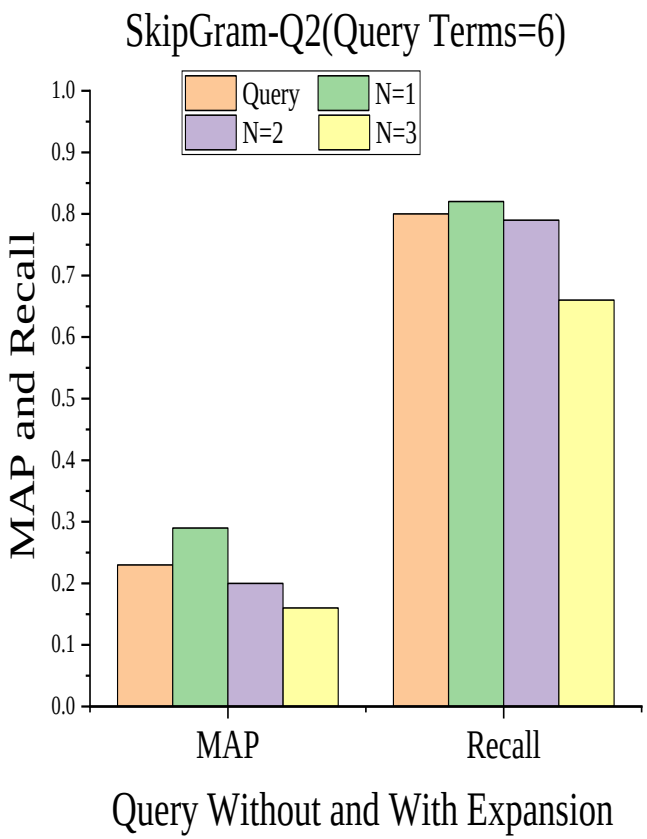
Expansion Using Skip-Gram

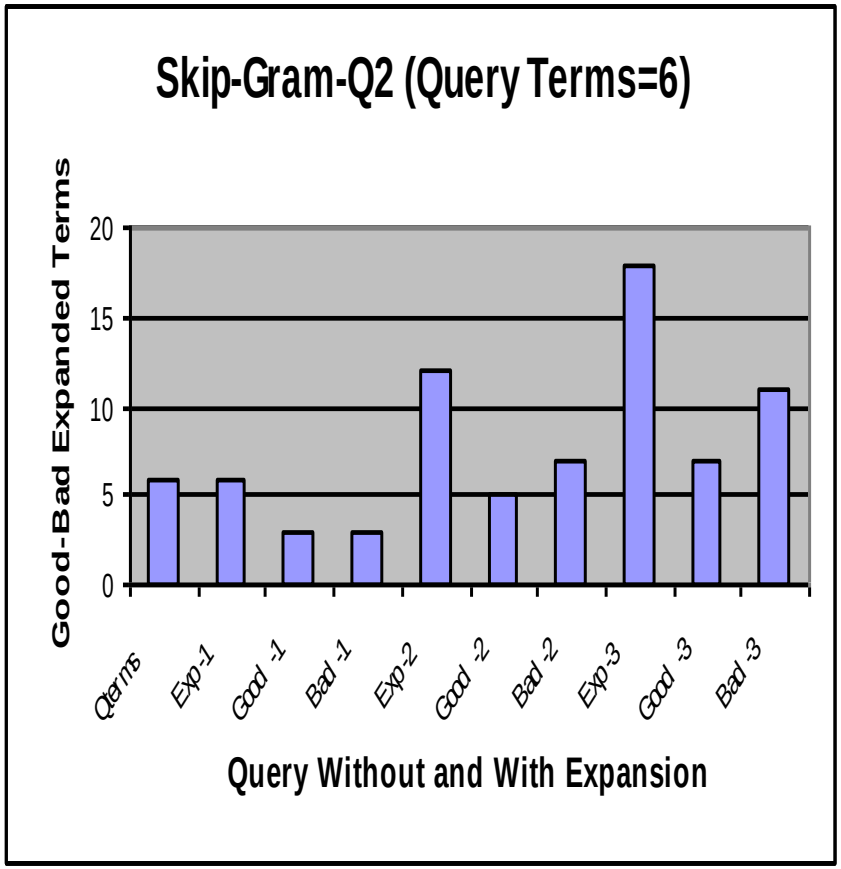




❑ Implementation and Experimental Results (cont.)

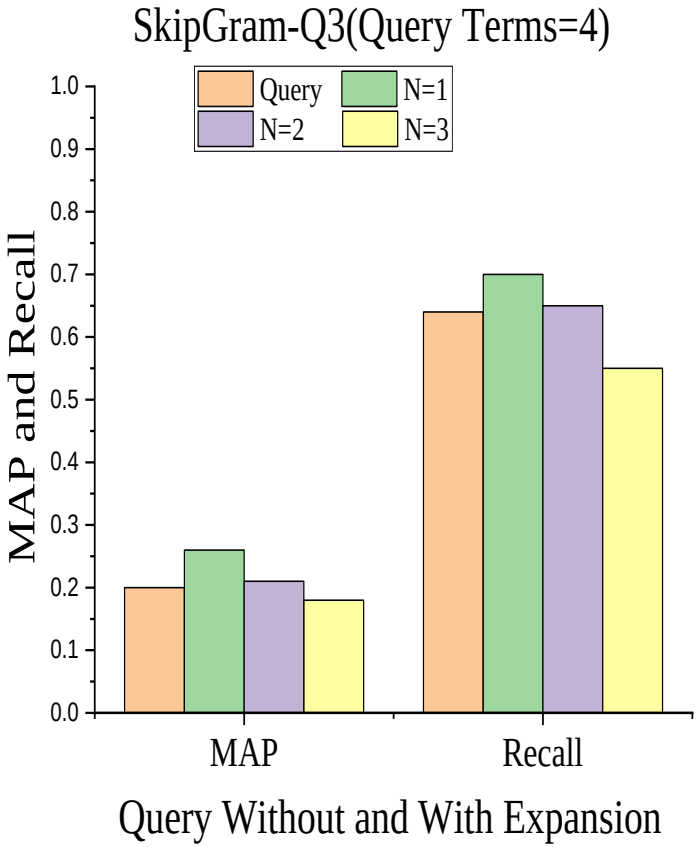
➢ Expansion Using Skip-Gram (cont.)

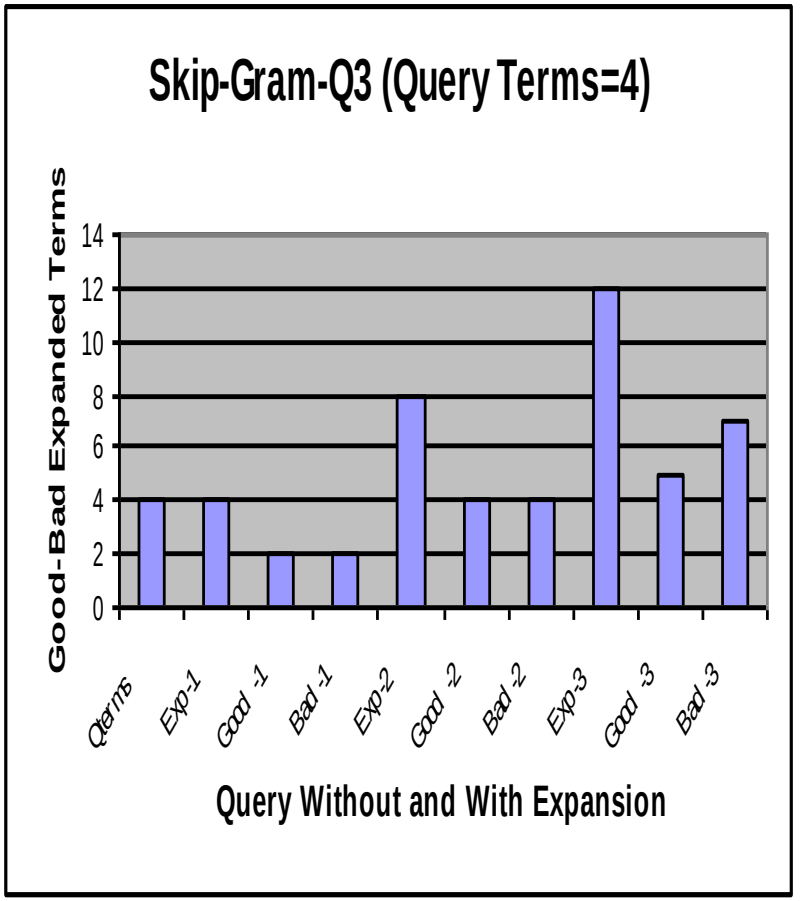




❑ Implementation and Experimental Results (cont.)

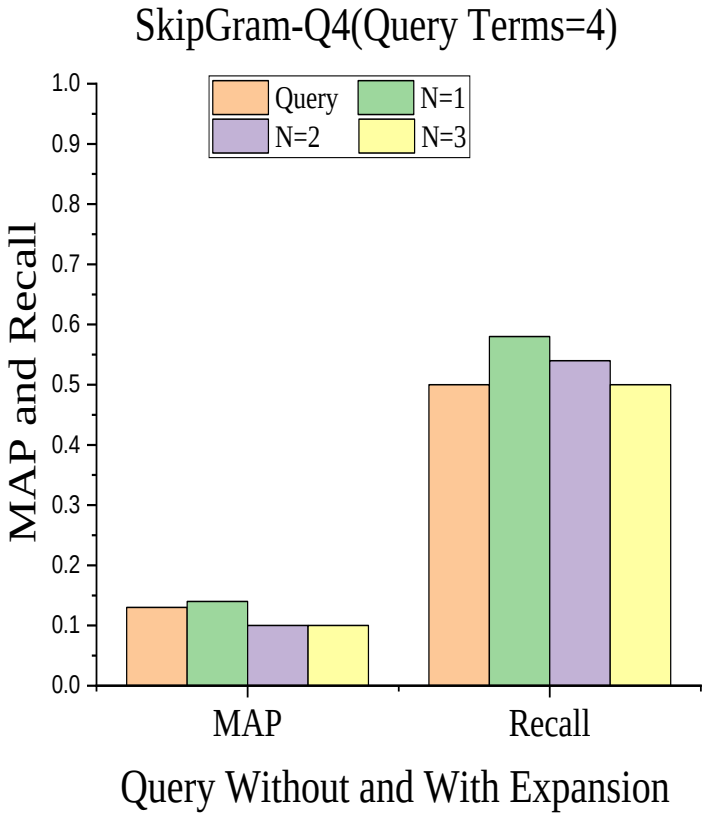
➢ Expansion Using Skip-Gram (cont.)

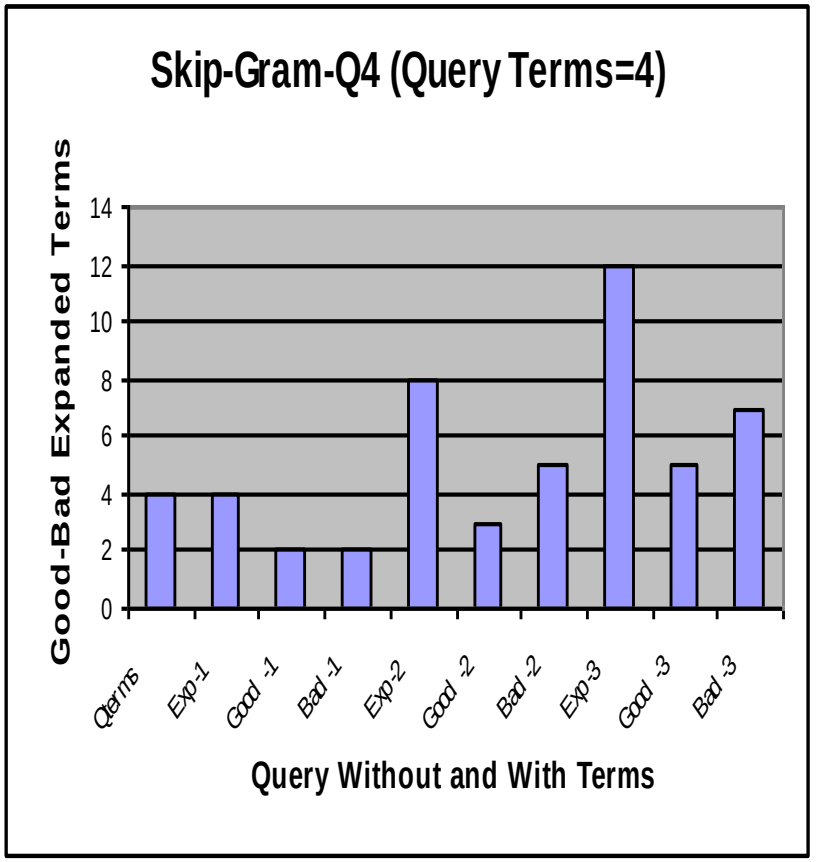




❑ Implementation and Experimental Results (cont.)

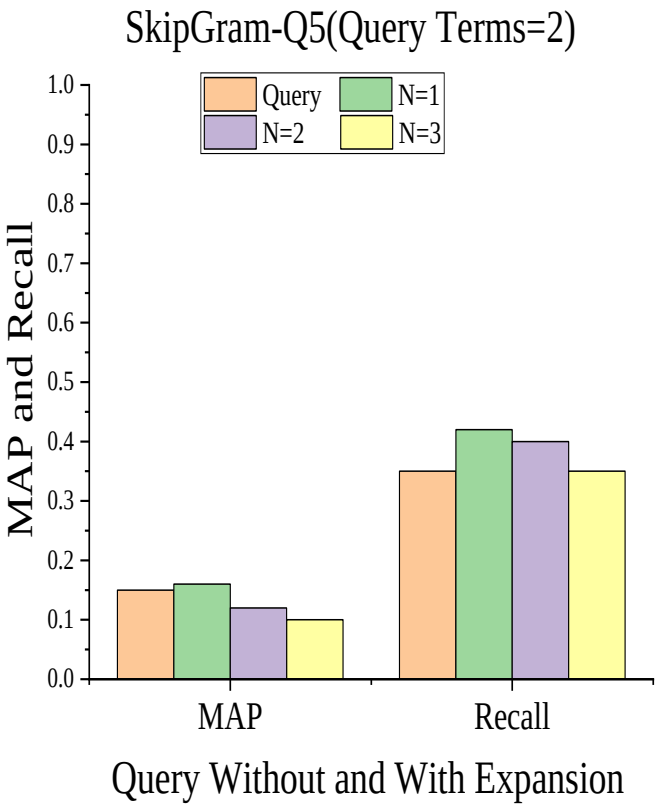
➢ Expansion Using Skip-Gram (cont.)

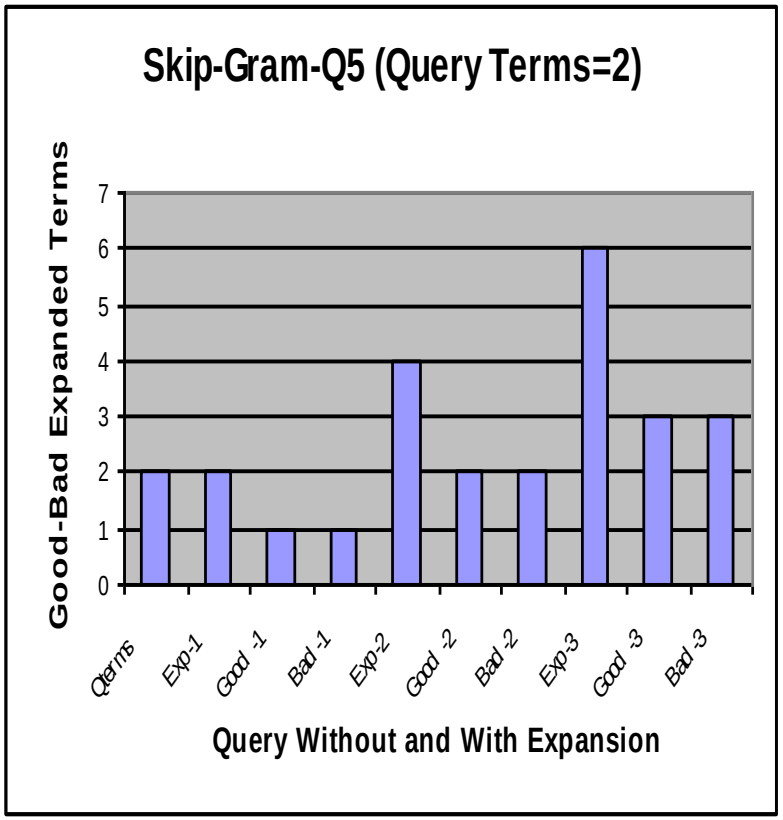




❑ Implementation and Experimental Results (cont.)

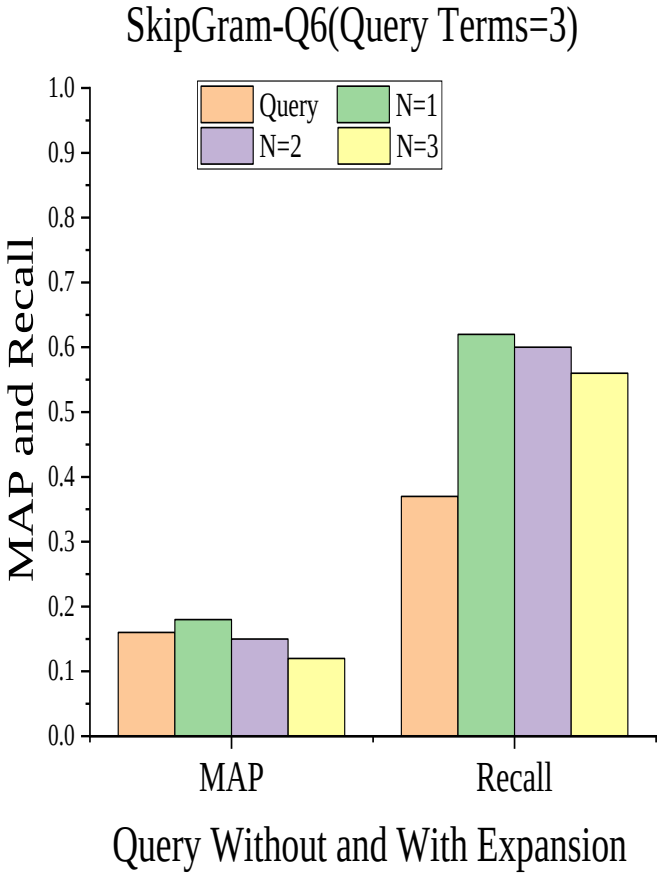
➢ Expansion Using Skip-Gram (cont.)

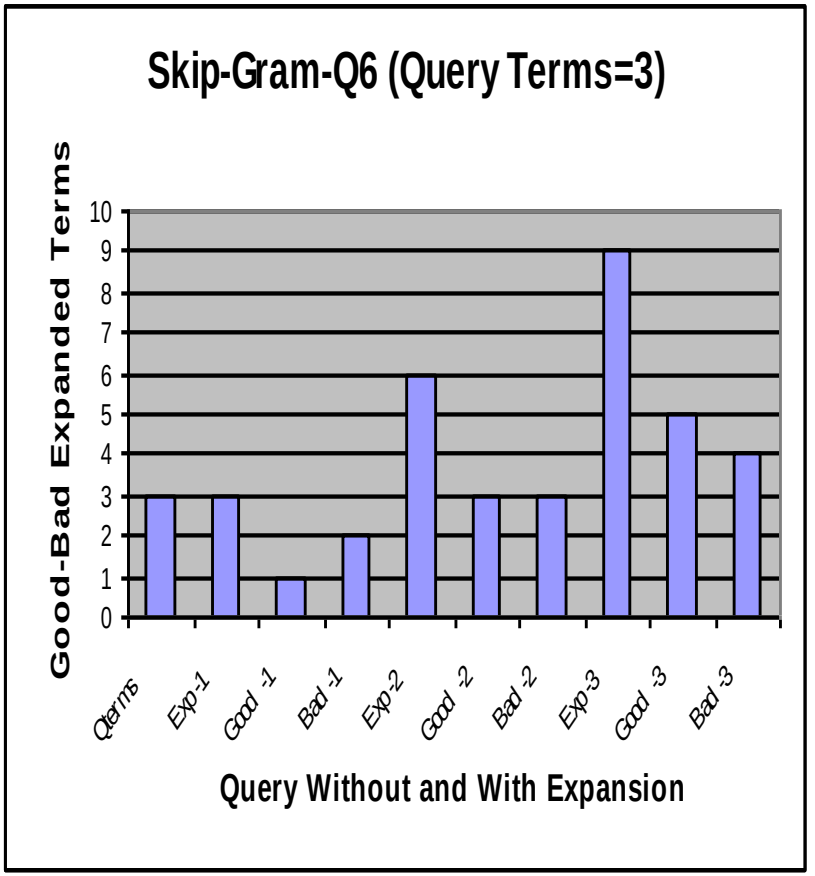




❑ Implementation and Experimental Results (cont.)

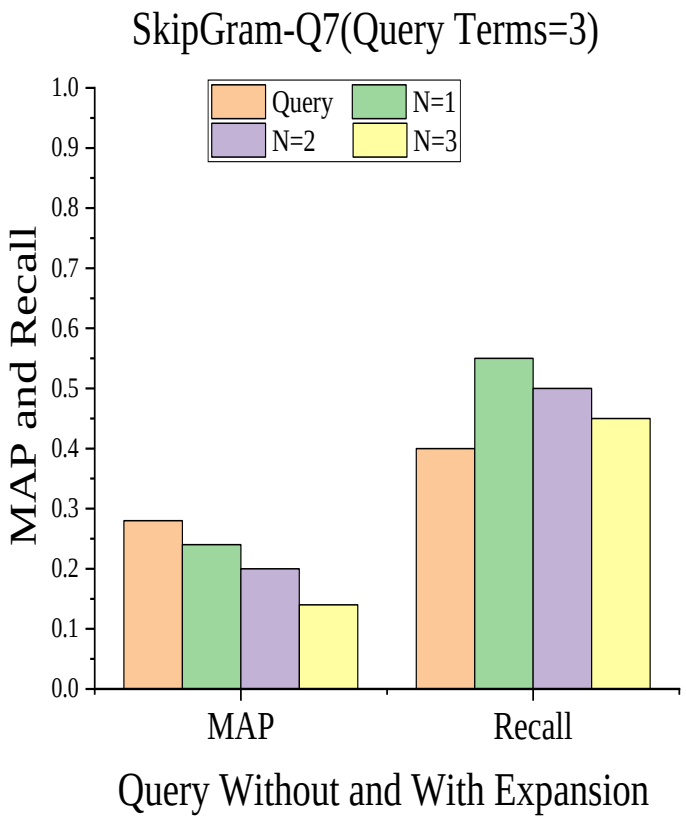
➢ Expansion Using Skip-Gram (cont.)

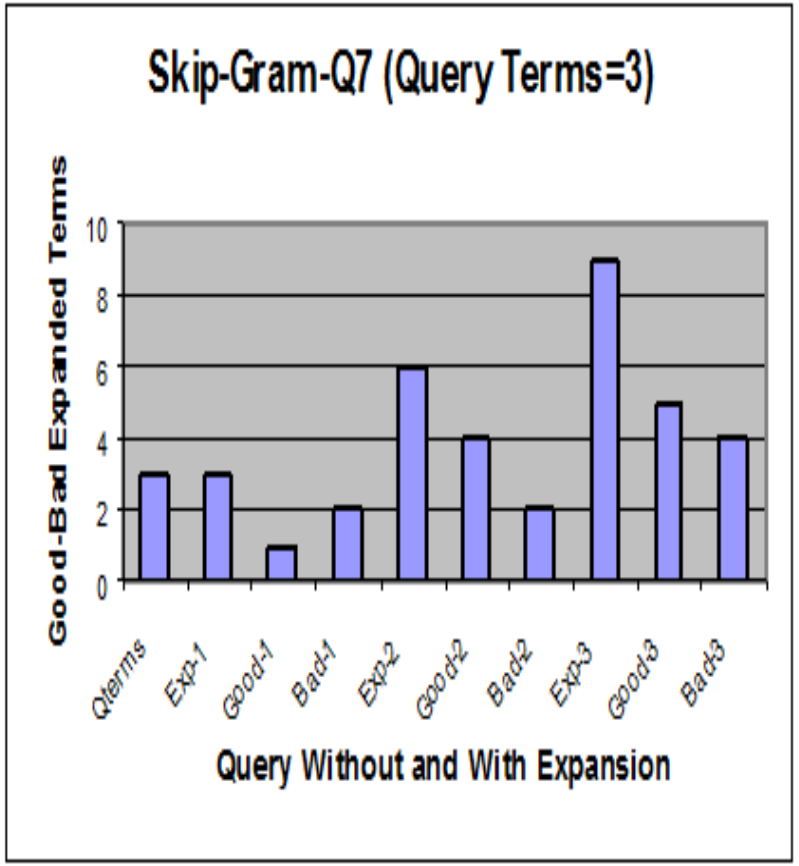




❑ Implementation and Experimental Results (cont.)

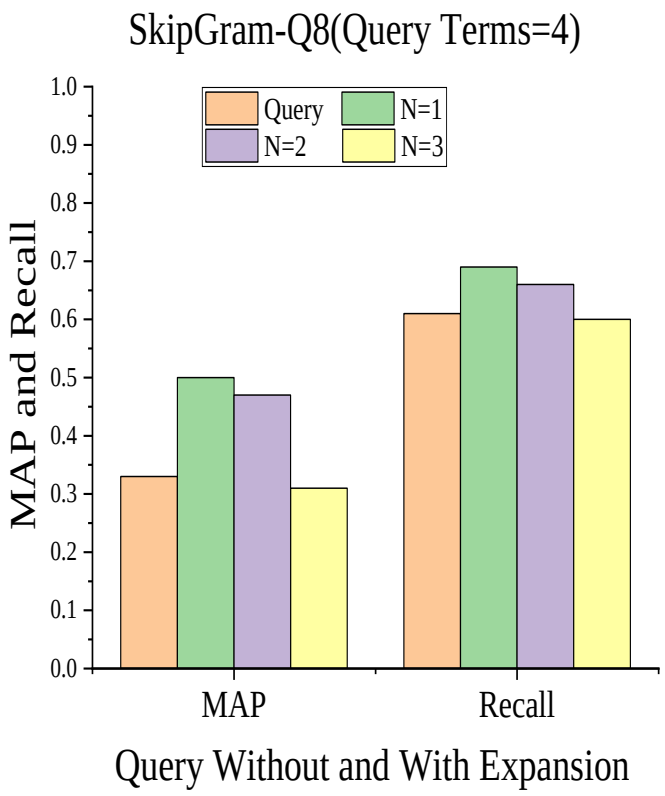
➢ Expansion Using Skip-Gram (cont.)

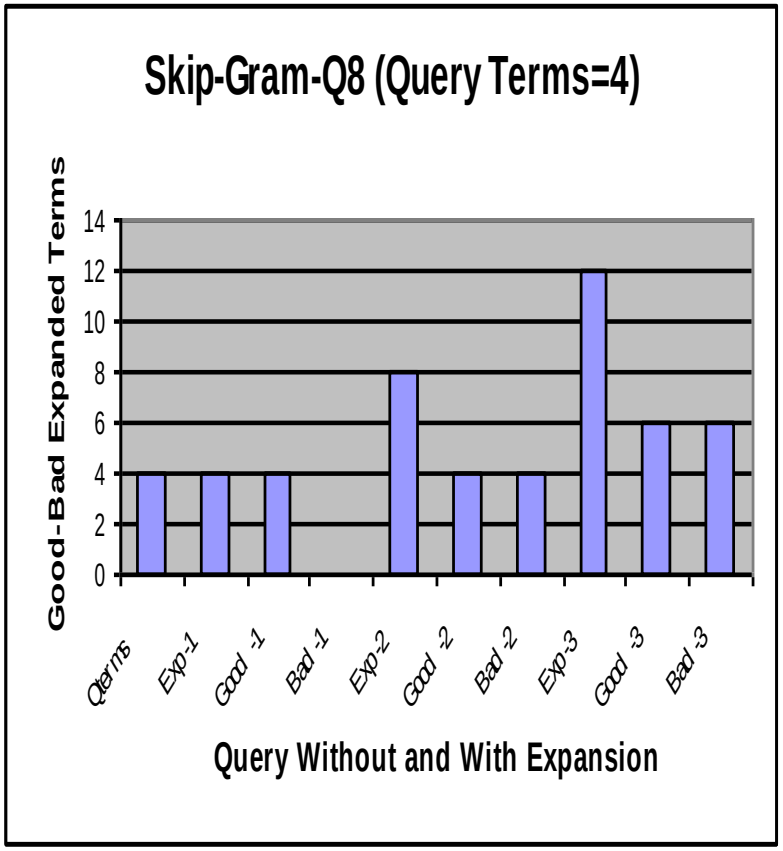




❑ Implementation and Experimental Results (cont.)

➢ Expansion Using Skip-Gram (cont.)



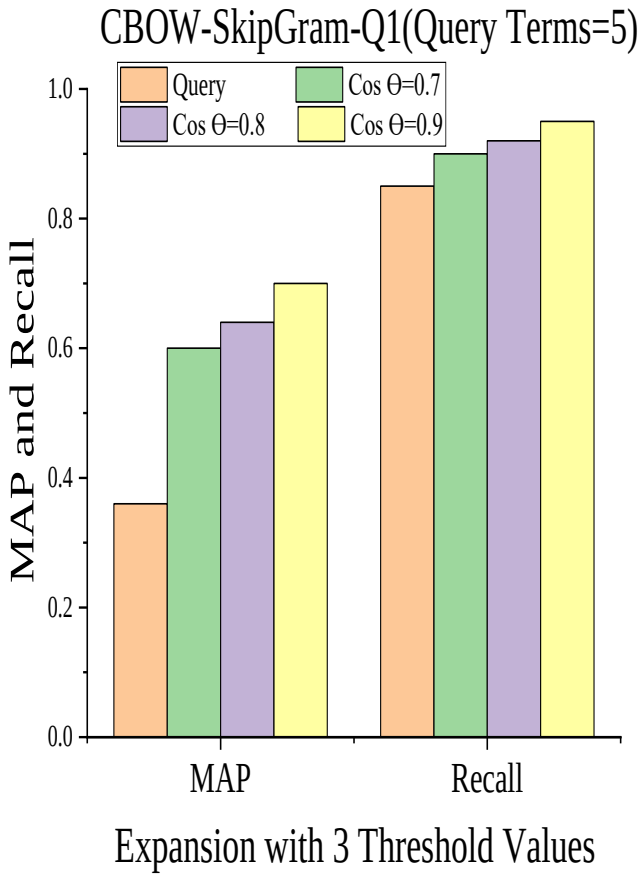


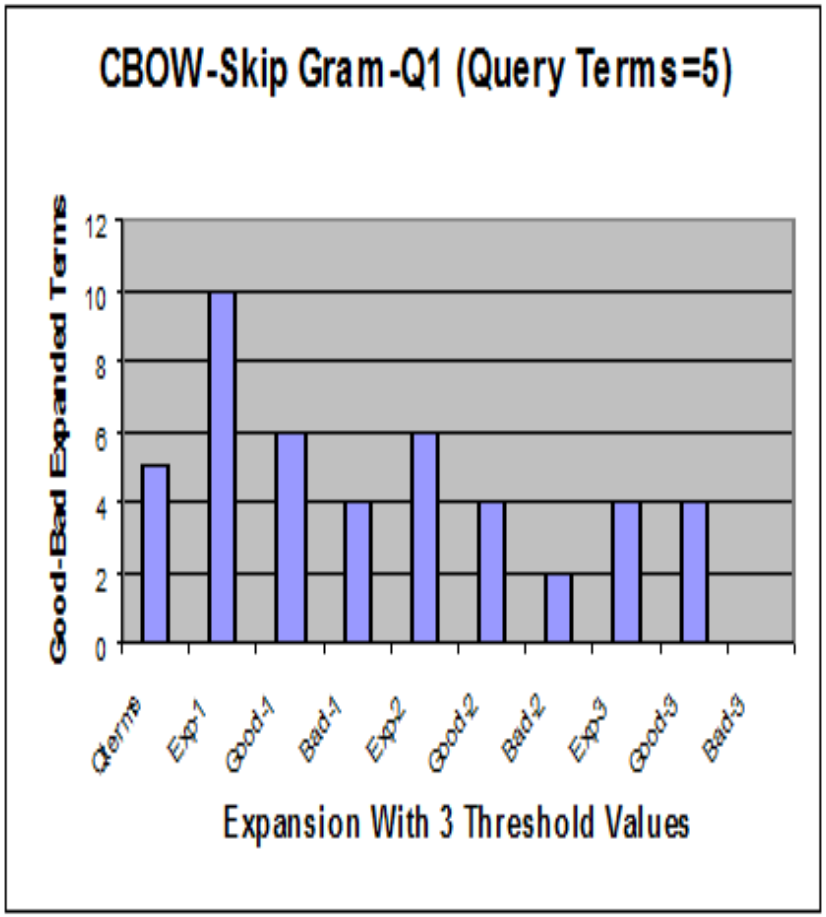
❑ Implementation of Hybrid Method

- The two approaches are combined and implemented to exploit the effectiveness of each.
- The similarity values between the query terms and those candidate ones are calculated.
- Changing the threshold value cos Ɵ (0.7, 0.8, 0.9).

❑ Implementation and Experimental Results (cont.)

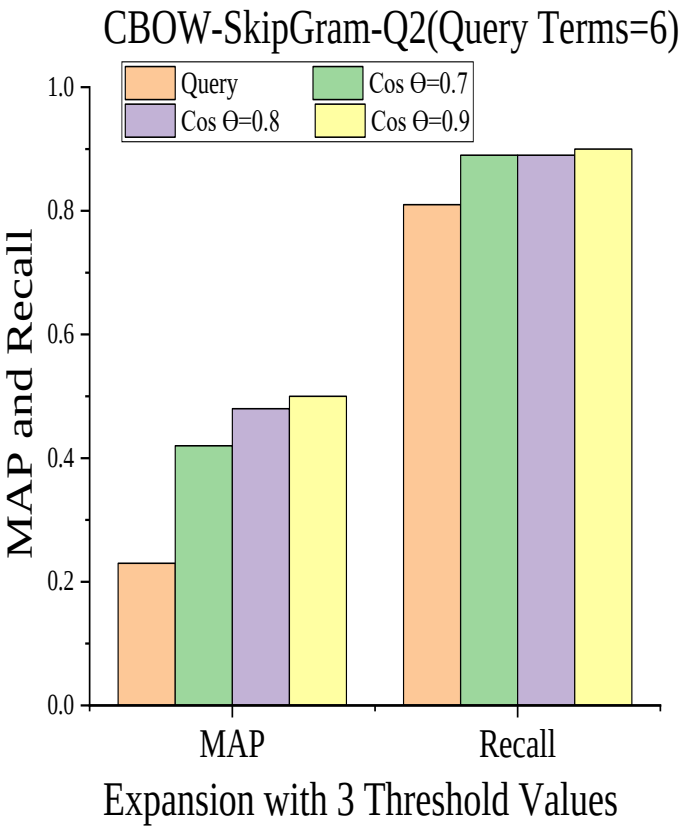
➢ Expansion Using CBOW and Skip-Gram

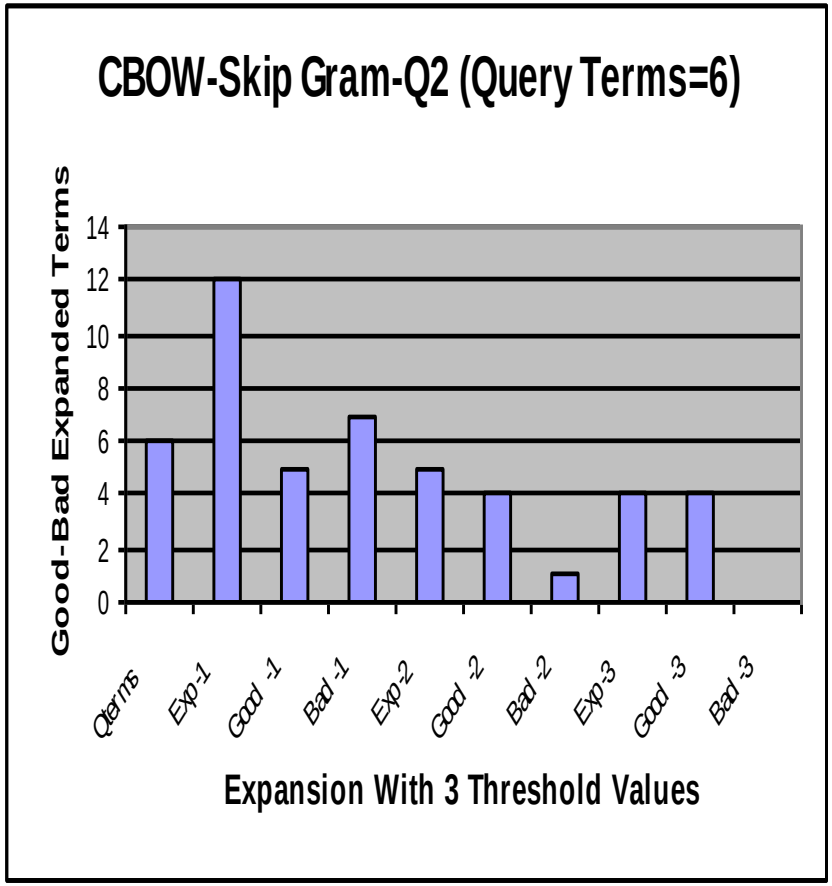




❑ Implementation and Experimental Results (cont.)

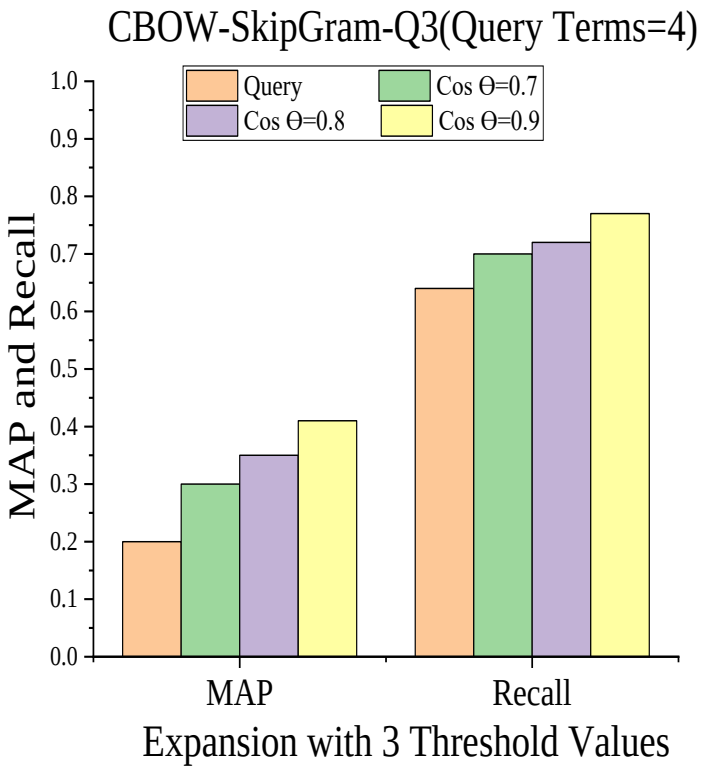
➢ Expansion Using CBOW and Skip-Gram (cont.)

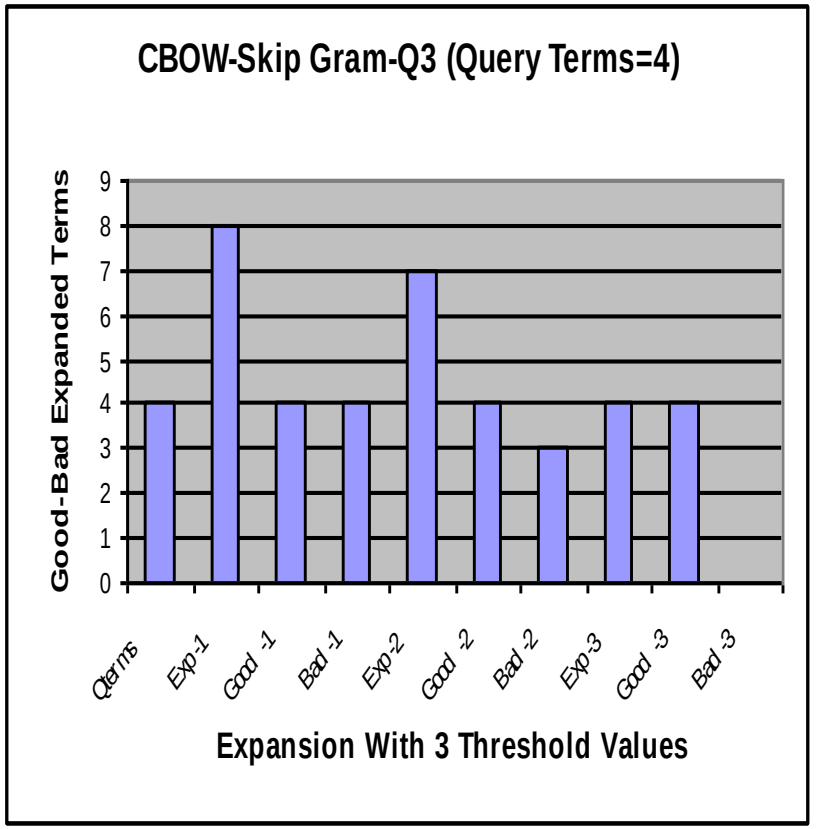




❑ Implementation and Experimental Results (cont.)

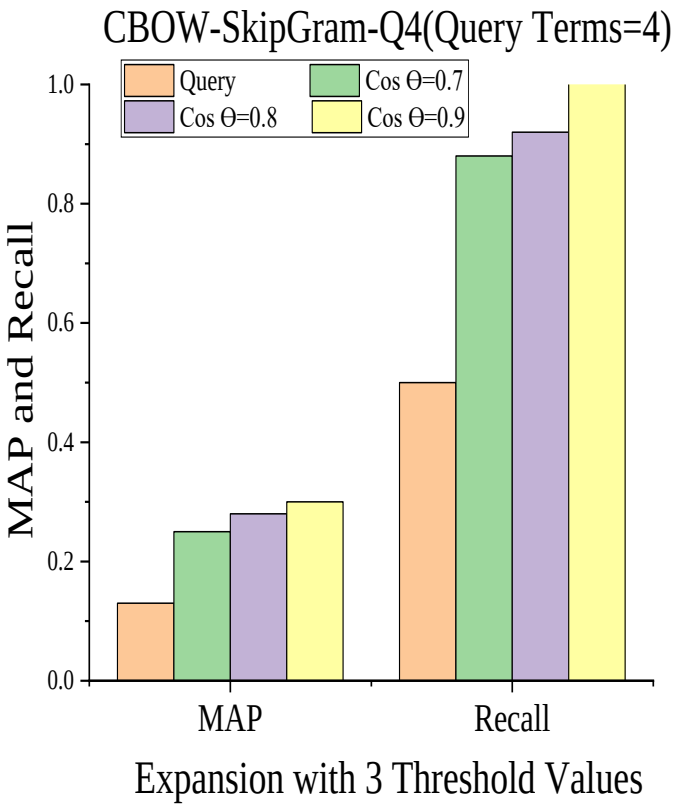
➢ Expansion Using CBOW and Skip-Gram (cont.)

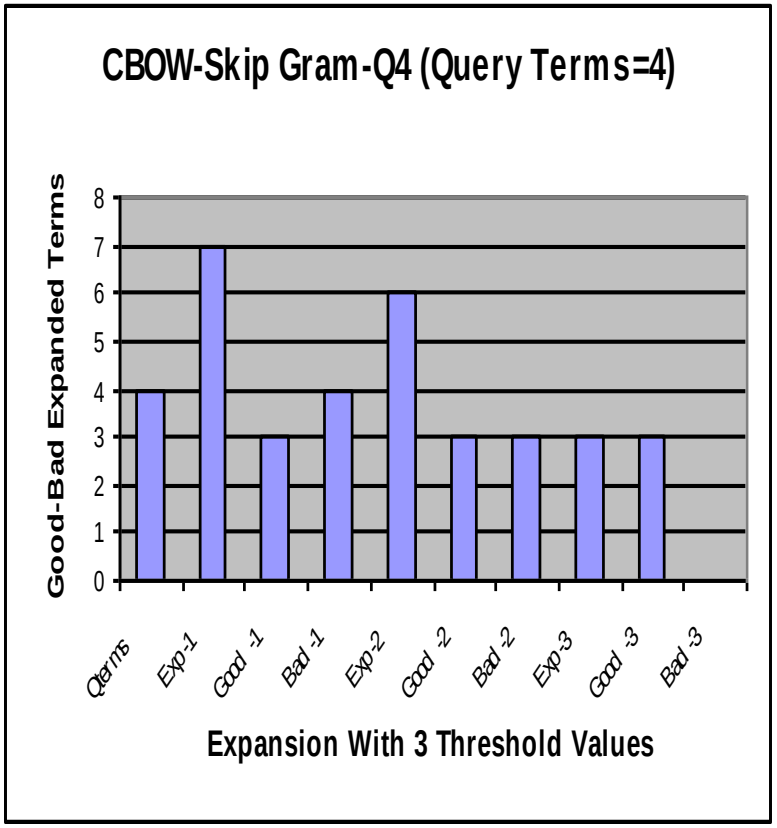




❑ Implementation and Experimental Results (cont.)

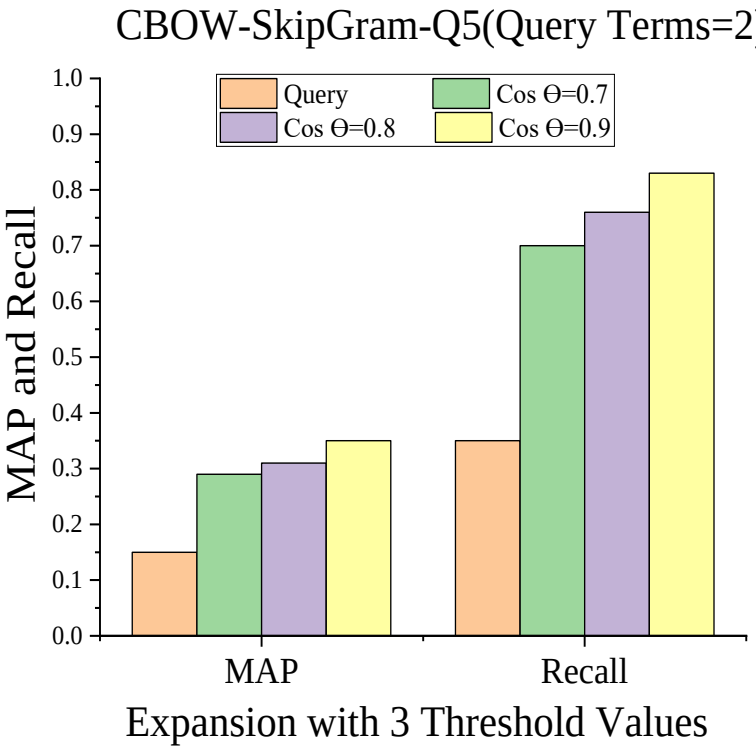
➢ Expansion Using CBOW and Skip-Gram (cont.)

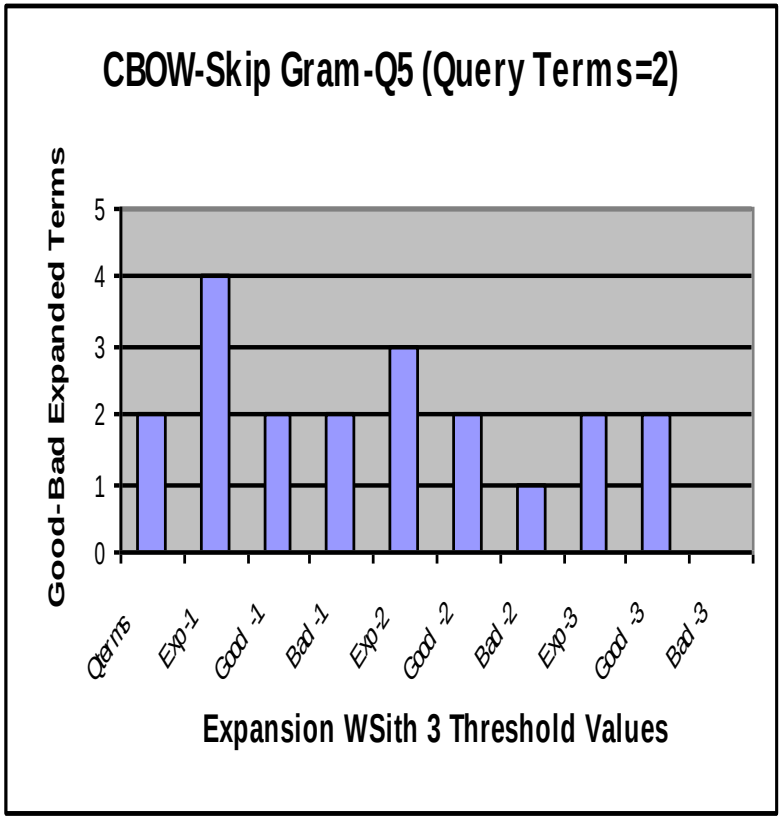




❑ Implementation and Experimental Results (cont.)

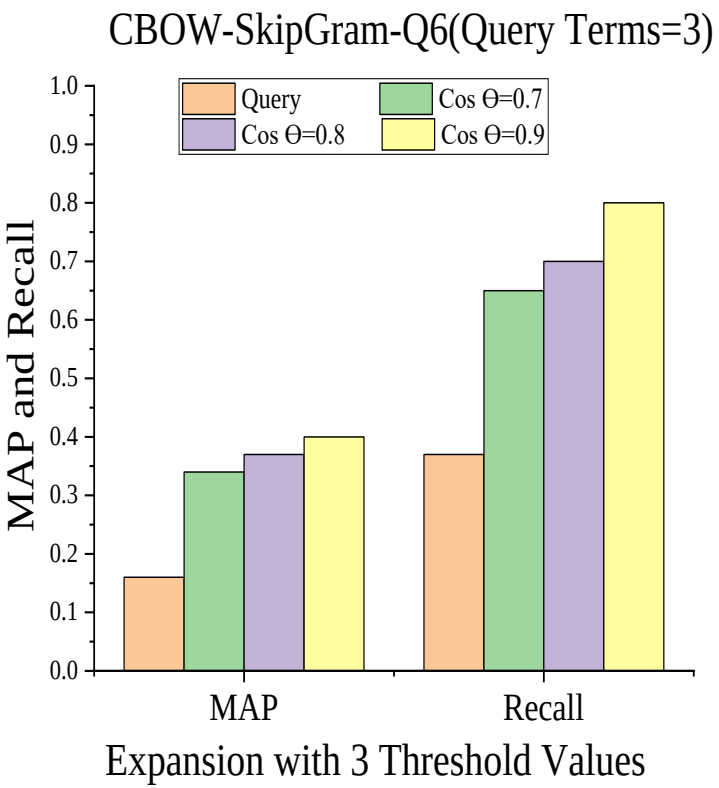
➢ Expansion Using CBOW and Skip-Gram (cont.)

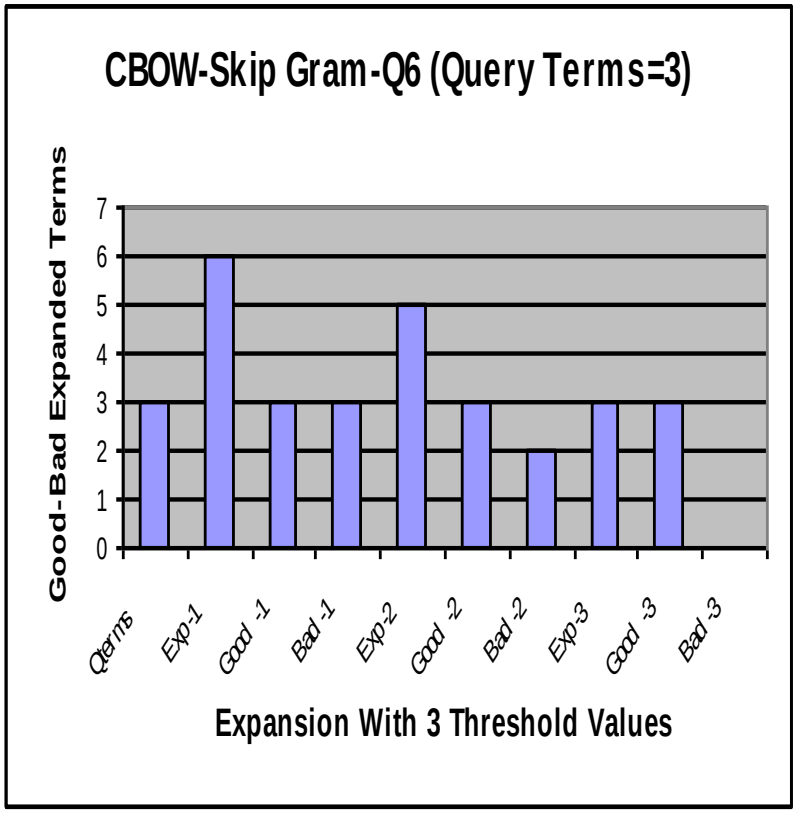




❑ Implementation and Experimental Results (cont.)

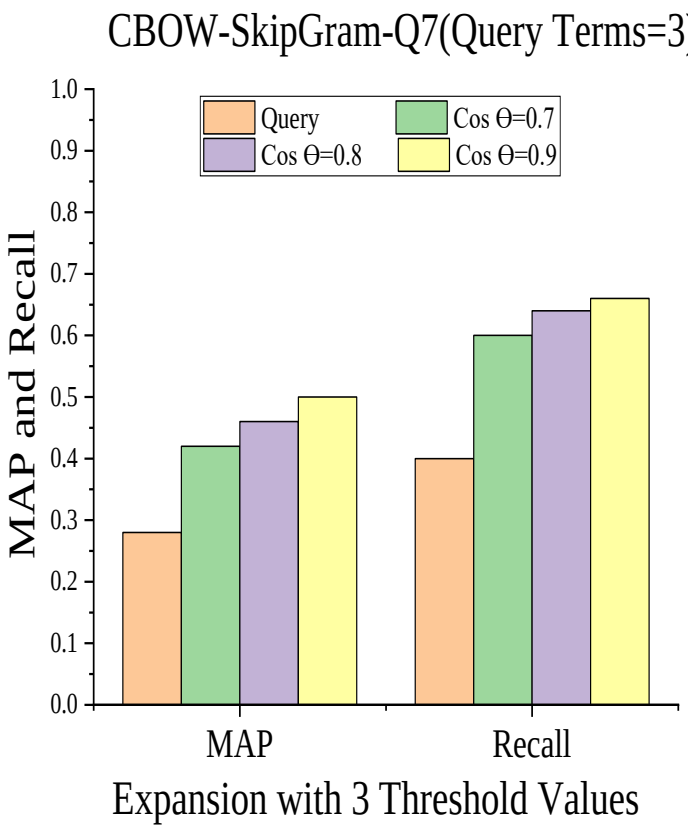
➢ Expansion Using CBOW and Skip-Gram (cont.)

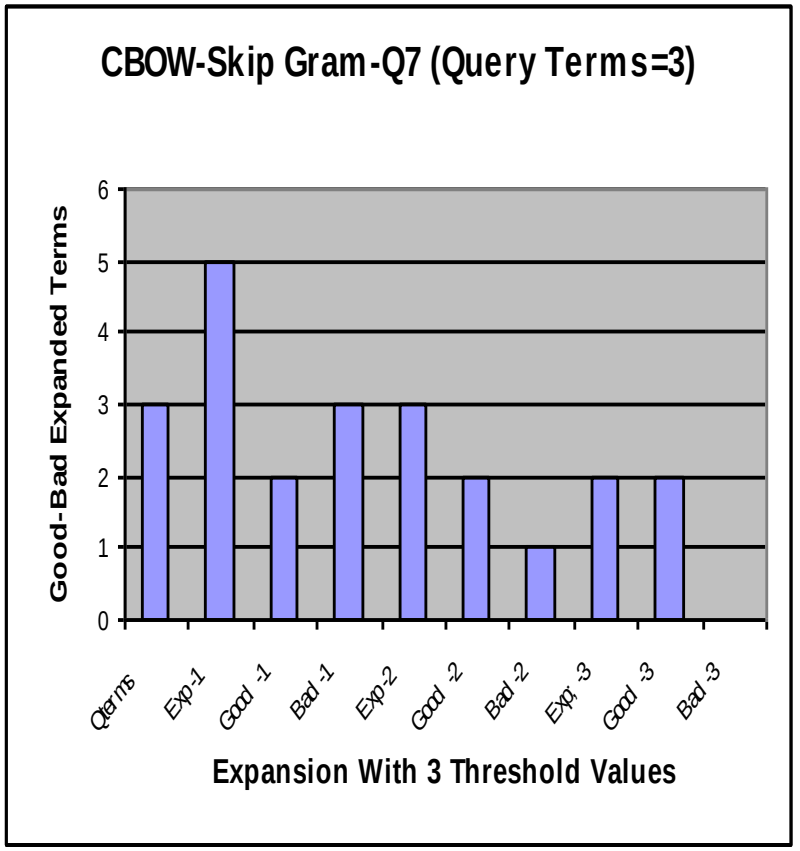




❑ Implementation and Experimental Results (cont.)

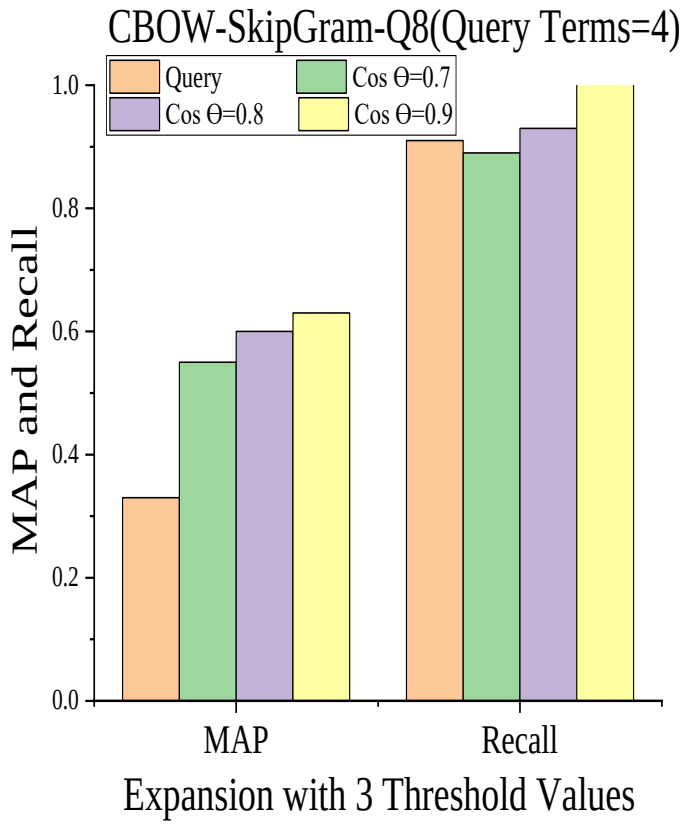
➢ Expansion Using CBOW and Skip-Gram (cont.)

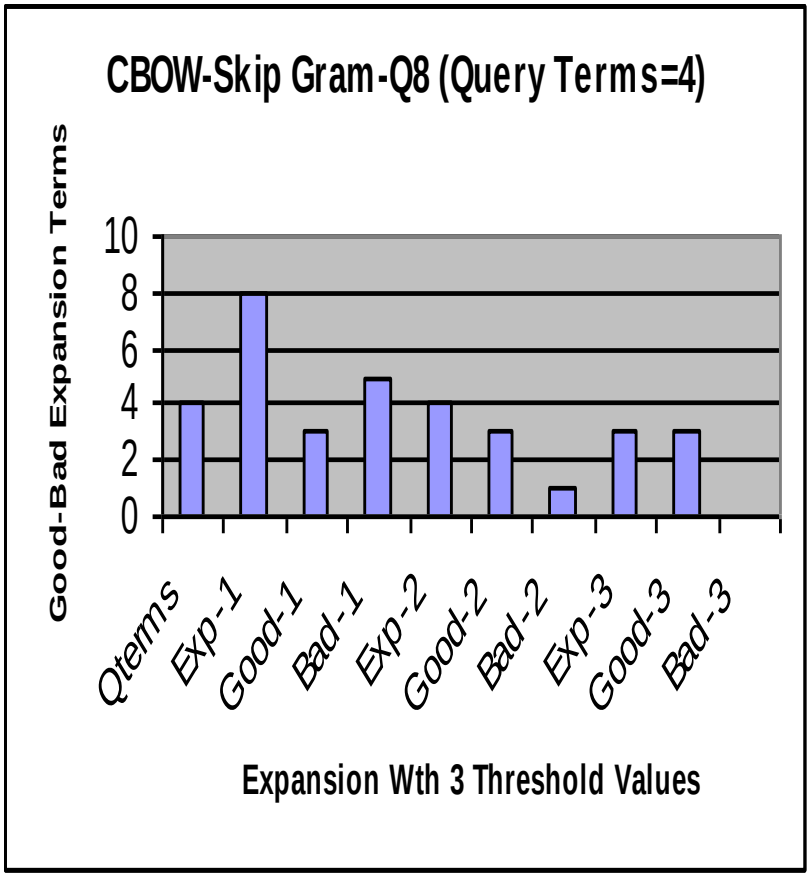




❑ Implementation and Experimental Results (cont.)

➢ Expansion Using CBOW and Skip-Gram (cont.)





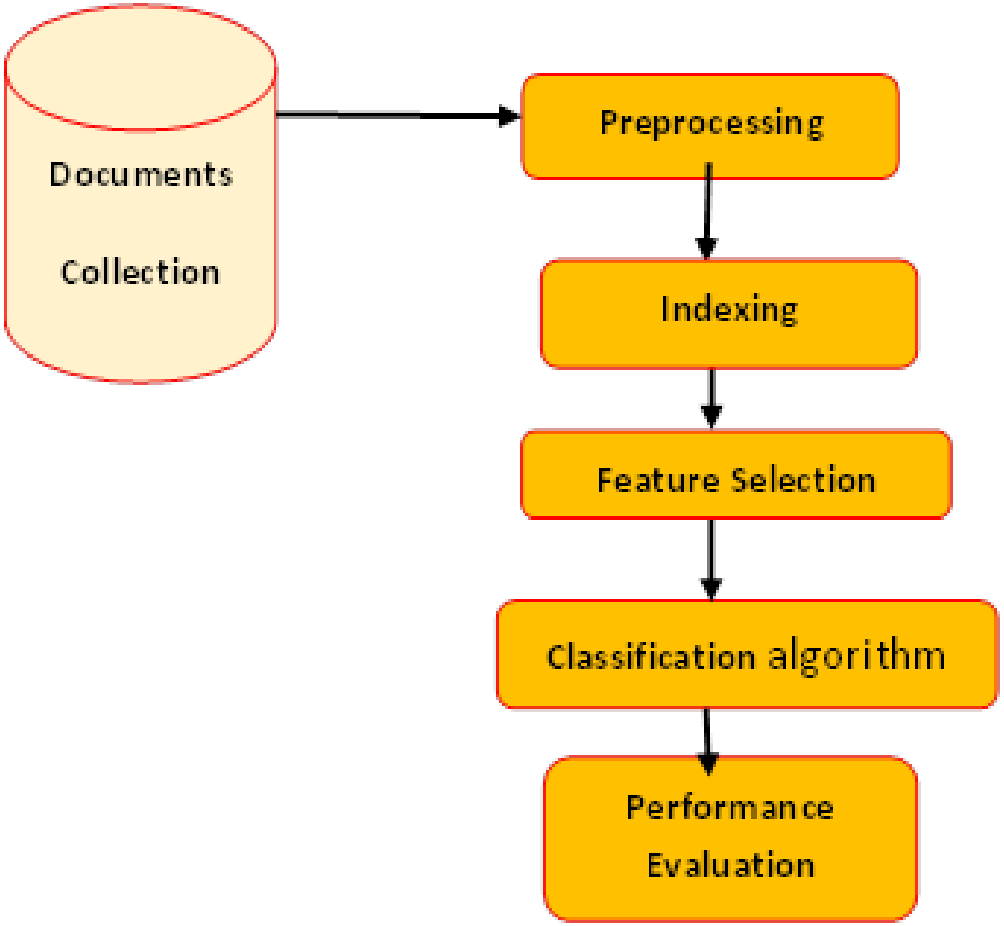
❑ Text Classification

- **Text classification** defined as assigning document or text to predefined categories or classes based on their content.

- One of the main problems of classifying documents is the huge number of features which are describing a dataset.
  - A huge number of features; in most cases; may reduce the efficiency of the adopted classifiers and also consume more time.

❑ Introduction of Text Classification (cont.)

➢ The Process of Arabic Text Classification



❑ Text Classification (cont.)

➢ The Process of Arabic Text Classification (cont.)

- **Feature selection**

- The main goal of feature selection is to select the more suitable and/or most significant features of the original document.

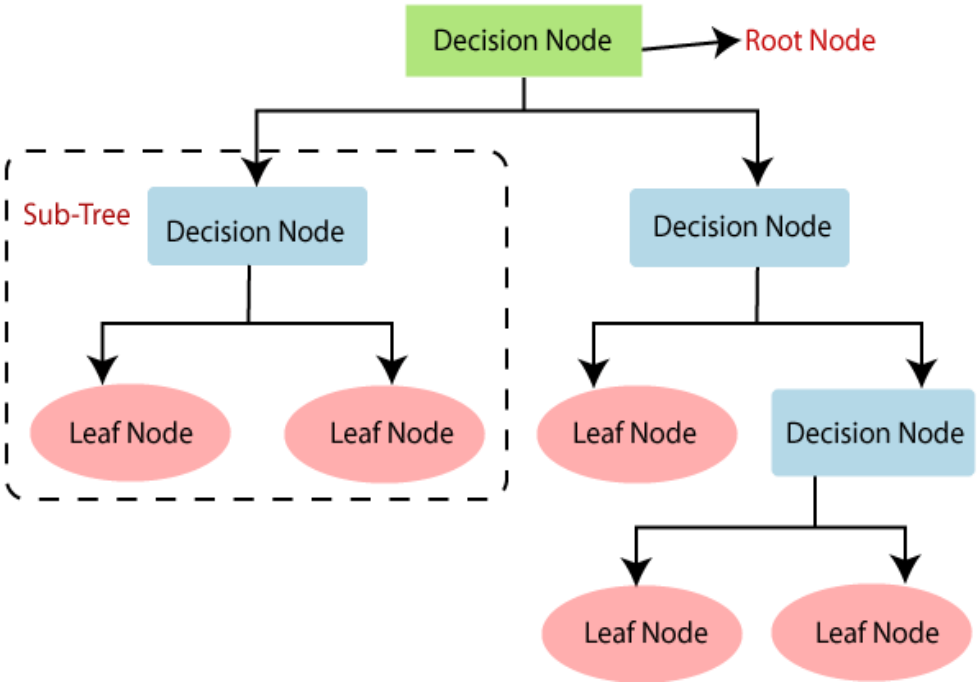
- **Classification Algorithm**

- The documents can be classified using the machine learning approaches such as Neural Network, Decision Tree, K-nearest neighbor (KNN), Naïve Bayes, Support Vector Machines (SVMs), and others.

❑ The Adopted Classification Algorithms

➢ Decision Trees

- Decision trees classifiers are prompt and simple data classifiers as supervised learning.



❑ The Adopted Classification Algorithms (cont.)

➢ Decision Trees (cont.)

- **Necessary Steps for Constructing tree are: -**

✓ Begin the tree with the root node, says S, which contains the complete dataset.

✓ Find the best attribute in the dataset using **Attribute Selection Measure (ASM).**

✓ Divide the S into subsets that contains possible values for the best attributes.

✓ Generate the decision tree node, which contains the best attribute.

✓ Recursively make new decision trees using the subsets of the dataset created in step -3. Continue this process until a stage is reached where you cannot further classify the nodes and called the final node as a leaf node.

❑ The Adopted Classification Algorithms (cont.)

➢ Naïve Bayes

- Naïve Bayes classifier is a simple probabilistic classifier based on applying Bayes' theorem.
- It used the following equation

$$P(\text{class} \mid \text{document}) = \frac{P(\text{class}).P(\text{document} \mid \text{class})}{P(\text{document})}$$

where

- P(class | document) is the probability that a given document D belongs to a given class C.
- P(document): The probability of a document.
- P(class) is the probability of a class (or category), it is computed as follows: $P(\text{class}) = \frac{\text{Number of documents in the category}}{\text{documents number in all categories}}$

❑ The Adopted Classification Algorithms (cont.)

➢ Naïve Bayes (cont.)

- ✓ P(document | class) represents the probability of a document given class.
- ✓ It is computed as follows: P(document | class) = Π p(word i | class).
- ✓ P(class | document) =p(class) Π p (word i |class).

Where

- ✓ P(word i | class) is the probability that the i-th word of a given document occurs in a document from class C, and this can be computed as follows:
- ✓ $P(\text{word i} \mid \text{class}) = (T_{ct} + \lambda) / (N_c + \lambda V)$

Where

- ✓ T_{ct}: The number of times the word occurs in that category C.
- ✓ N_c: The number of words in category C.
- ✓ V: the size of the vocabulary table.
- ✓ λ: the positive constant, usually 1, or 0.5 to avoid zero probability.

❑ The Adopted Classification Algorithms (cont.)

➢ Support Vector Machine (SVM)

- The main idea of SVM classifier is mapping the input points in N-dimension space into another higher dimensional space.

- Found a maximal separating hyper plane.

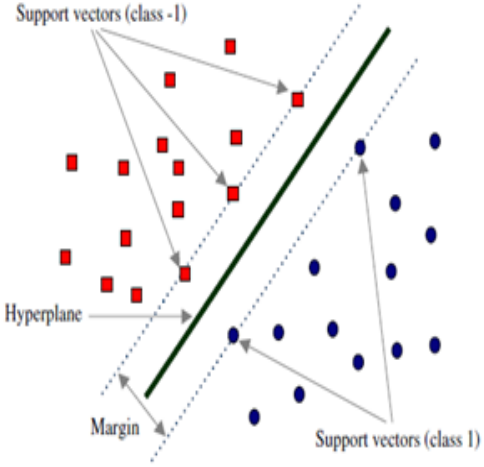
- It aims to separate amounts of data based on the optimal hyper plane between vectors which are linearity separable.

❑ The Adopted Classification Algorithms (cont.)

➢ SVM (cont.)

To obtain a nonlinear SVM, there are two main steps.

- The original input data is transformed into a higher dimensional space using a nonlinear mapping.
- Searches for a linear separating hyper plane in the new Space.



The linear algorithm of this classification process can be defined as:

$$f(x) = W.p + b \qquad (12)$$

Where

- P is the vector of the training data-set.
- b is the bias, to manipulate the decision boundary of the linear hyper plane.
- W is the weight vector for the best hyper-plane.

❑ Implementation of Three Classifiers

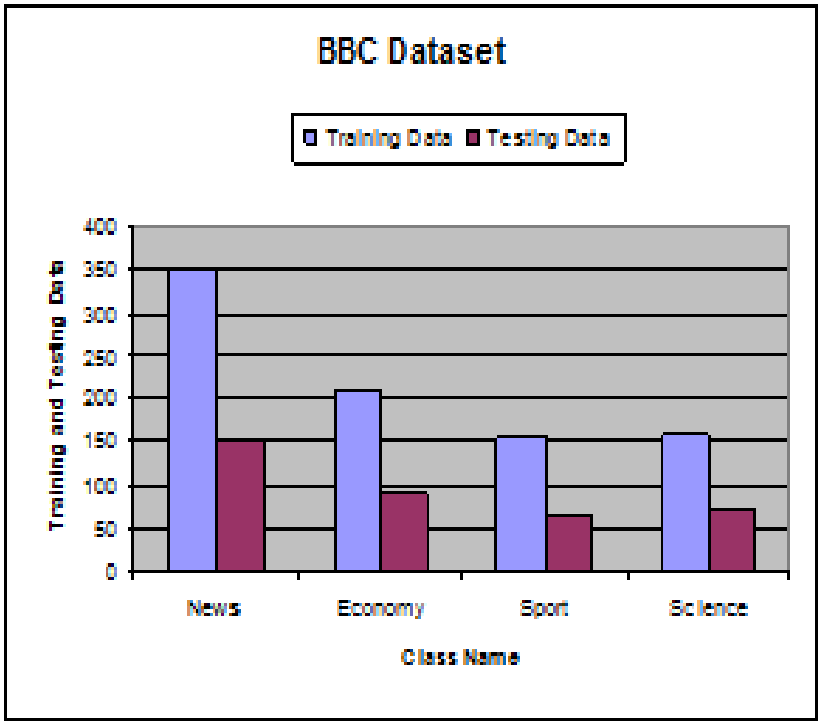
➢ Document Collection Dataset

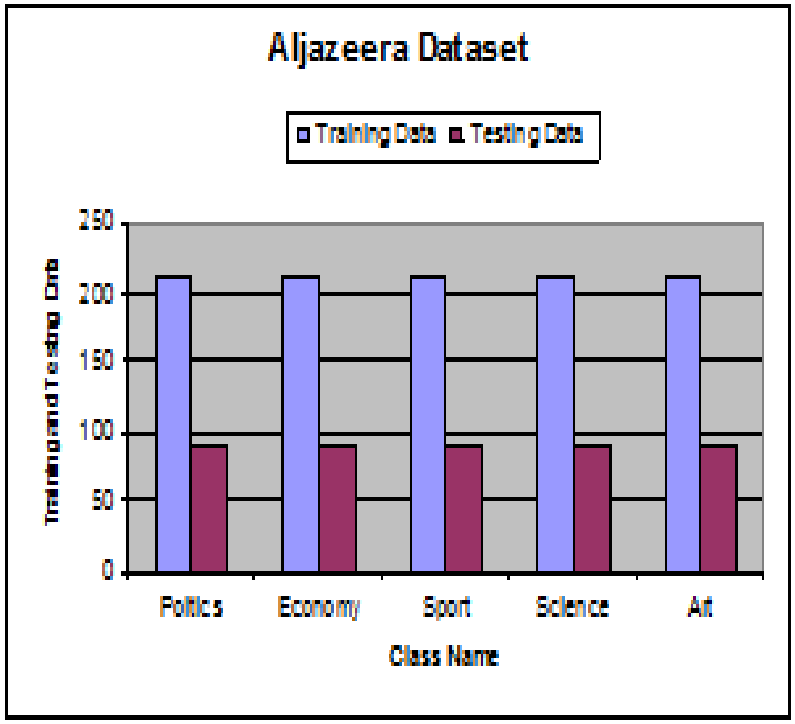
- BBC Dataset
  - It contained 1250 document in Arabic language.
  - It contains four categories {'World News', 'economy', 'sport', 'Science'}.
- Aljazeera Dataset
  - It contained 1500 document in Arabic language.
  - It contains five categories {'Politics', 'Science', 'Sport', 'Economy', 'Art'}.

The dataset is split as follows: 70% for training and the remaining 30% for testing.

❑ Implementation of Three Classifiers

➢ Document Collection Dataset (cont.)





❑ Implementation of Three Classifiers (cont.)

➢ Performance Metrics

- **Recall:** is the probability that a randomly selected relevant document D is retrieved in a search.

$$Recall = \frac{TP}{TP+FN}$$

- **Precision**: is the probability that a randomly selected retrieved document D is relevant to the predicted class C.

$$Precision = \frac{TP}{TP + FP}$$

- **F- Measure**: It can be determined from Precision and Recall which is the harmonic mean of precision and recall.

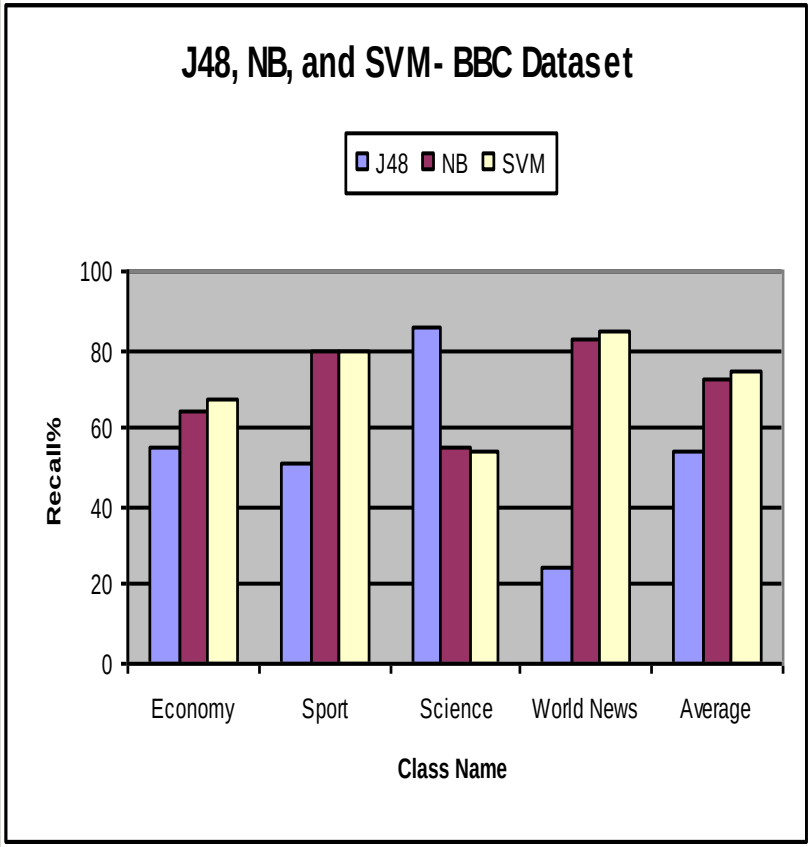
$$F - measure = \frac{2*TP}{2*TP+FP+FN}$$

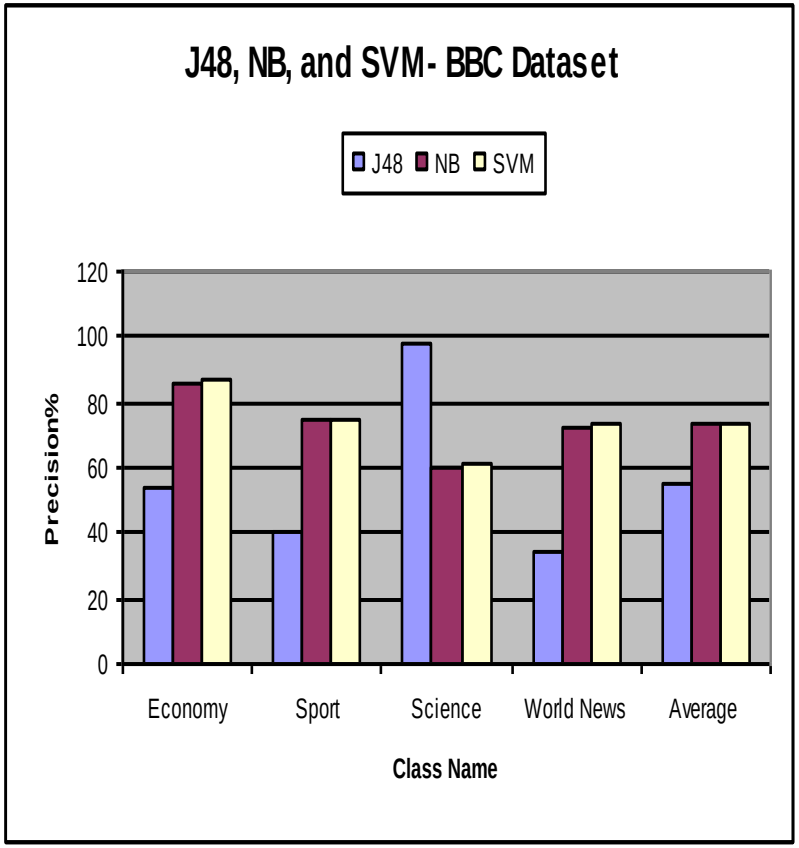
Where

- ✓ TP is the True Positive
- ✓ FN is the False Negative
- ✓ FP is the False Positive

❑ Experimental Results

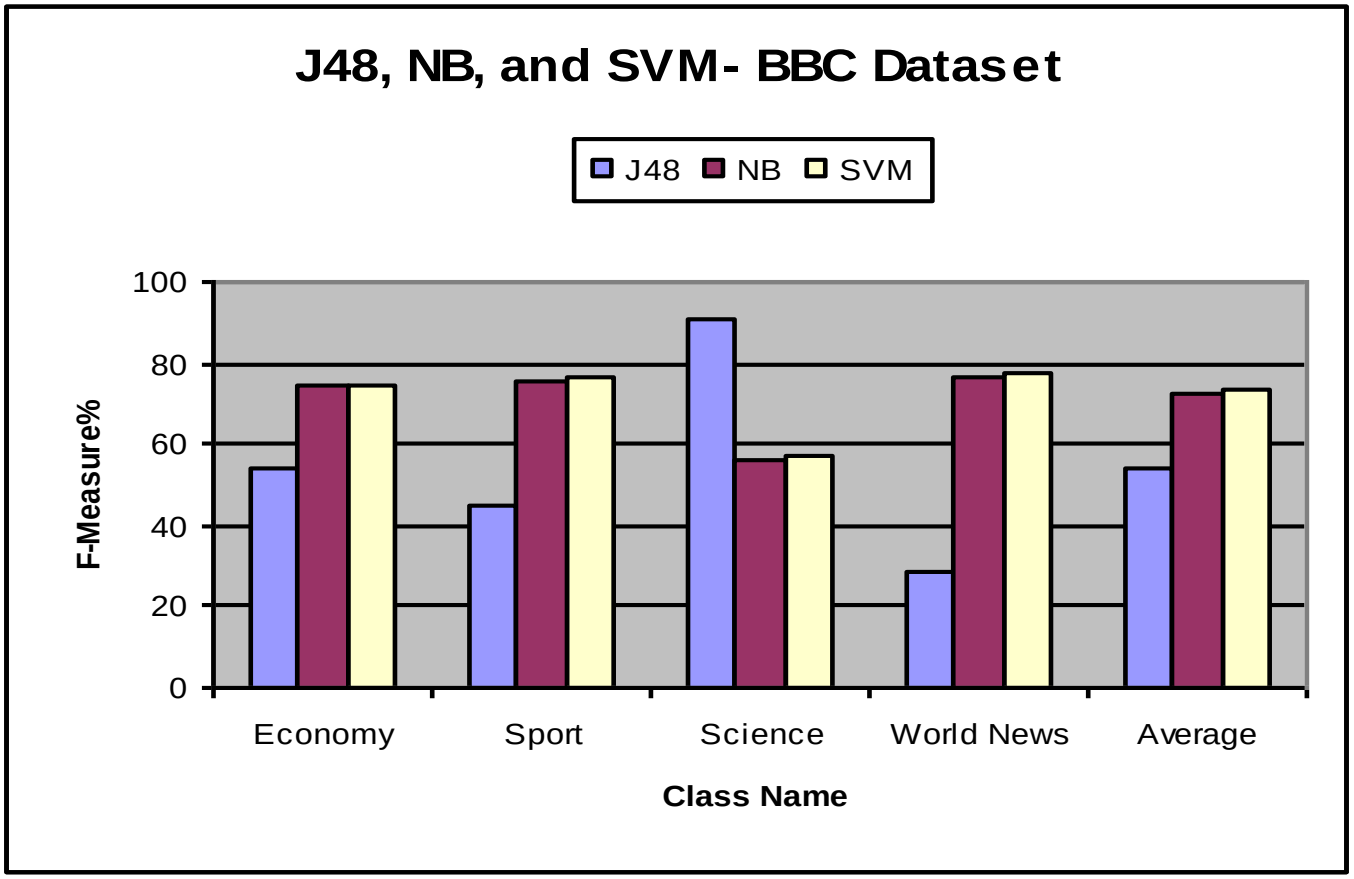
➢ Classifiers Operated on BBC Dataset





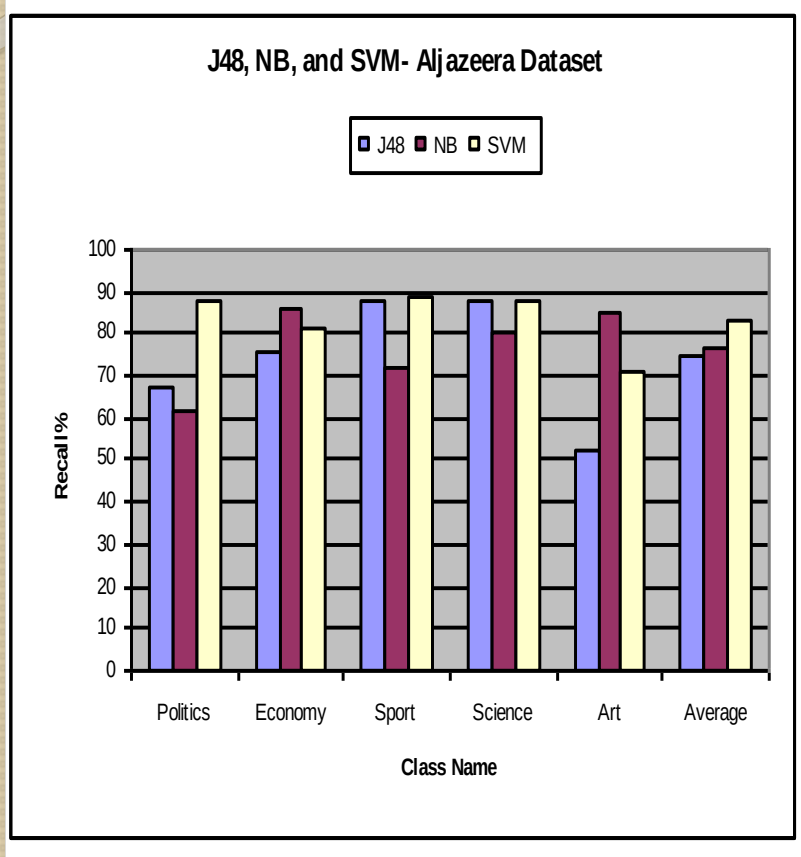
❑ Experimental Results (cont.)

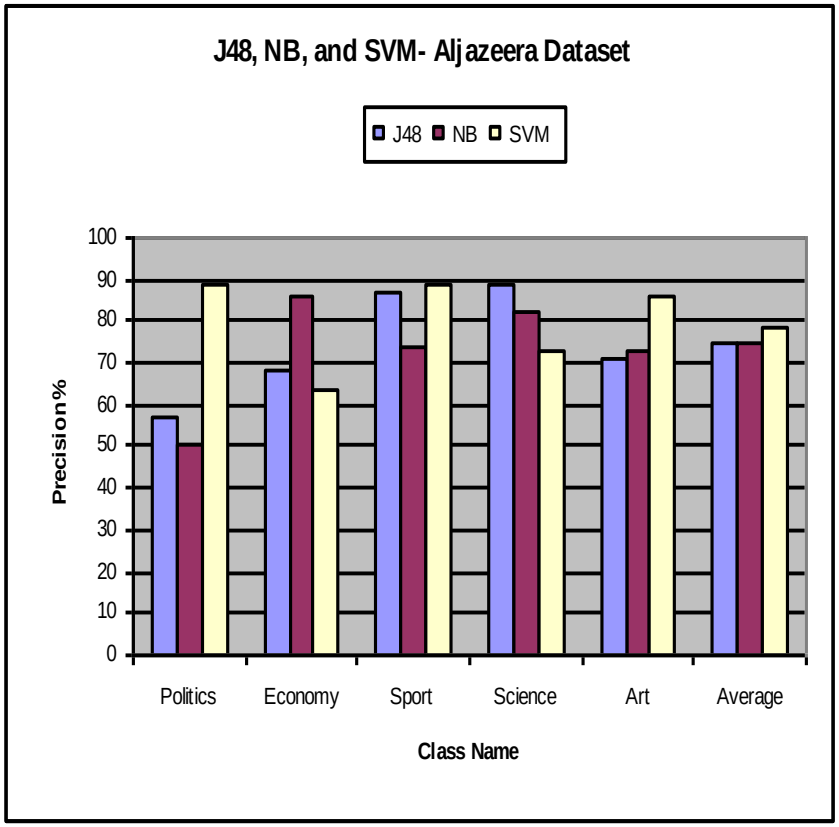
➢ Classifiers Operated on BBC Dataset (cont.)



❑ Experimental Results (cont.)

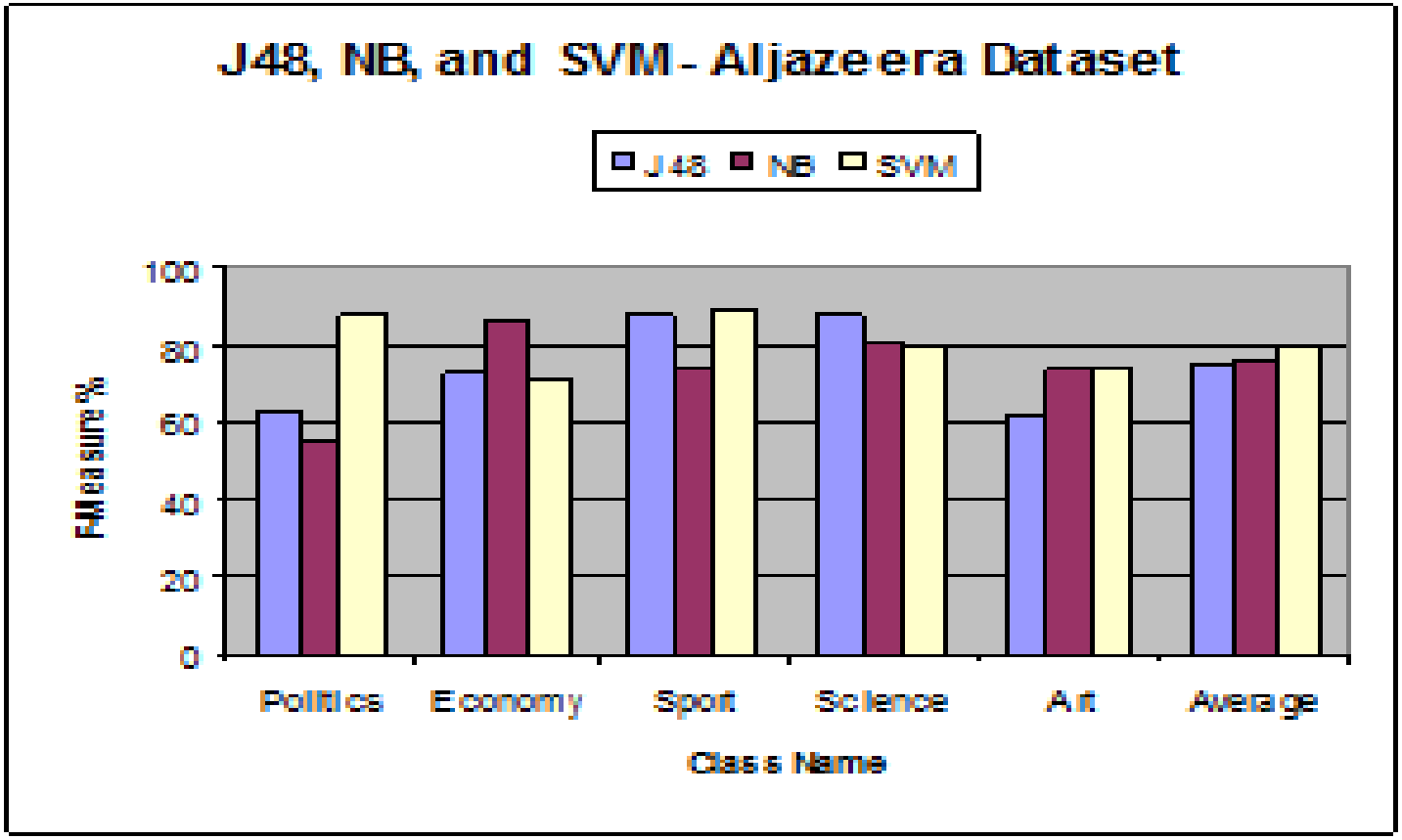
➢ Classifiers Operated on Aljazeera Dataset





❑ Experimental Results (cont.)

➢ Classifiers Operated on Aljazeera Dataset (cont.)



❑ Feature Selection Methods

➢ Term Weighting (TF-IDF)

- It is used to calculate the term weighting.

- Each document d_i is represented as a vector of terms weights $w_{i,j}$ as follows:

$$d_i = (w_{i,1}, w_{i,2}, \ldots, w_{i,j}, \ldots, w_{i,t})$$

- In each document, the term weighting is assigned for each term according to the term frequency.

❑ Feature Selection Methods

➢ Term Weighting (TF-IDF)

- The term weighting is calculated by

$$w_{i\,j} = tf(i,j) * idf(i,j) = tf(i,j) * \log(n/df(j))$$ where

✓ $w_{i,j}$ represents the weight of term j in document i

✓ tf (i,j) represents the frequency of term j in a document i

✓ idf (i,j) is a factor used to improve the term which has low frequency and appears in a few documents.

$$idf(i,j) = \log(n/df(j))$$

Where

✓ n is the number of all documents in the dataset

✓ df (j) is the number of documents which contains the term j.

❑ Feature Selection Methods (cont.)

➢ Chi-square

- Chi-square is a nonparametric statistical filter method that is used to compute the lack of independence between the distributions of observed frequencies and the theoretically expected frequencies.
- The value of the chi-square statistic is given by:

where $$x^2 = \sum_{i=1}^{m} \sum_{j=1}^{k} \frac{\left(A_{ij} \frac{R_i . c_j}{N}\right)^2}{\frac{R_i . c_j}{N}}$$

m is the number of intervals

- ✓ k the number of classes
- ✓ N the total number of samples
- ✓ R_i the number of patterns in the i interval
- ✓ c_j the number of patterns in the j class
- ✓ A_{ij} the number of patterns in the i interval and the j class.

❑ Feature Selection Methods (cont.)

➢ Gini Index (GI)

- It can be considered as an improved version of the attribute selection method used in the construction of decision tree.
- This measure is defined as follows:

$$GI(t) = \sum_{i=1}^{M} p(t|C_i)^2 p(C_i|t)^2$$

where

✓ p $(t|C_i)$ is the probability of term t given class C_i

✓ p $(C_i|t)$ is the probability of C_i in the presence of t.

❑ Feature Selection Methods (cont.)

➢ Information Gain (IG)

- IG measure can suggest the importance of the features, by calculating the weight (relevance) of a feature in terms of the class features.

- If feature is higher weight, this feature is better.

❑ Feature Selection Methods (cont.)

➢ Information Gain (cont.)

- IG of a feature f is mathematically expressed as follows:

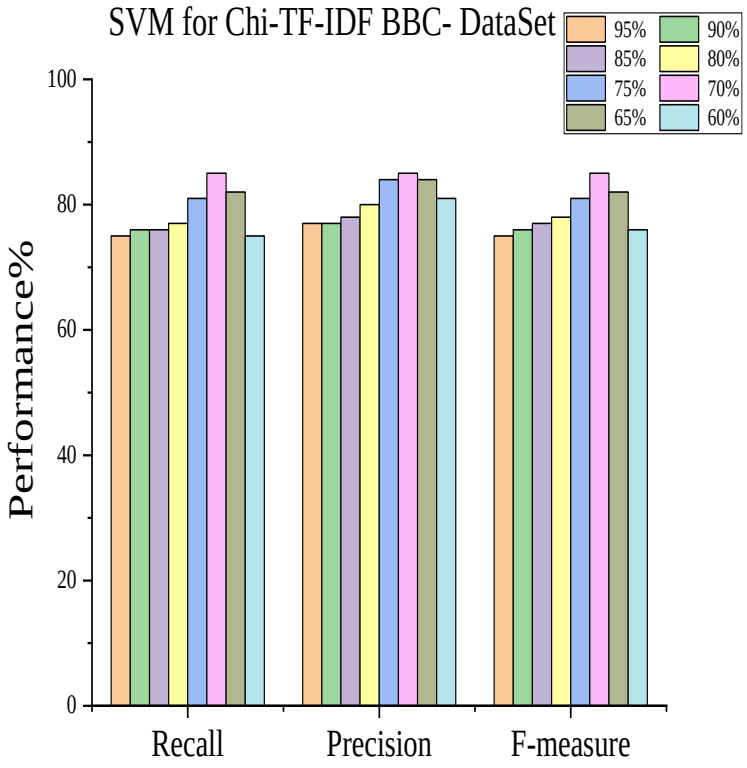
$$IG = -\sum_{i=1}^{M} p(c_i)\log p(c_i) + p(t)\sum_{i=1}^{M} p(c_i|t)\log p(c_i|t) + p(\bar{t})\sum_{i=1}^{M} p(c_i|\bar{t})\log p(c_i|\bar{t}) \quad (16)$$

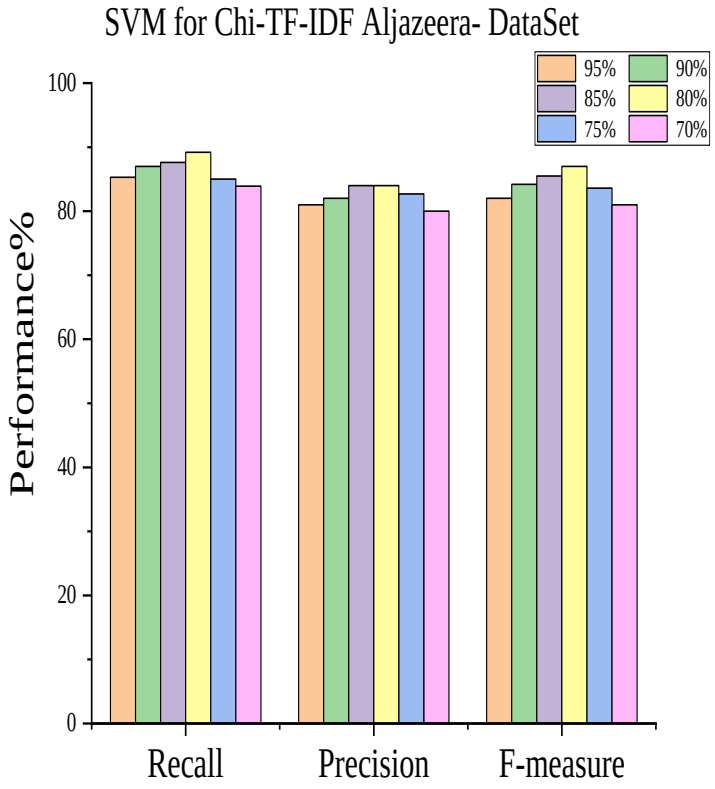
where

- ✓ M is the number of categories
- ✓ $p(c_i)$ is the probability of category c_i
- ✓ p(t) is the probability of occurrence,
- ✓ $p(\bar{t})$ are the probabilities of nonappearance of feature t,
- ✓ $p(c_i|t)$ and $p(c_i|\bar{t})$ are the conditional probabilities of category c_i considering the presence and absence of feature t, respectively.

❑ Experimental results

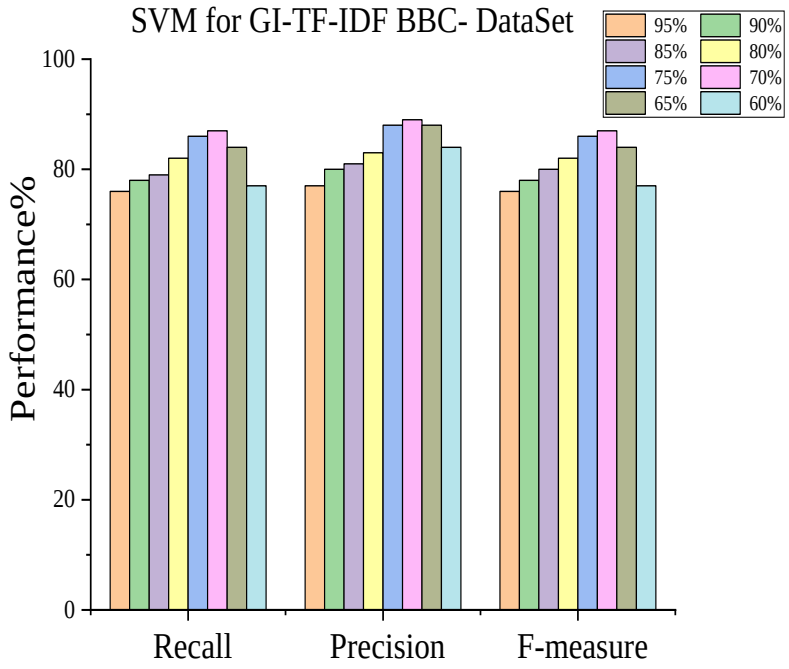
➢ Enhancement Using Combination of Chi Square and TF-IDF Feature Selection

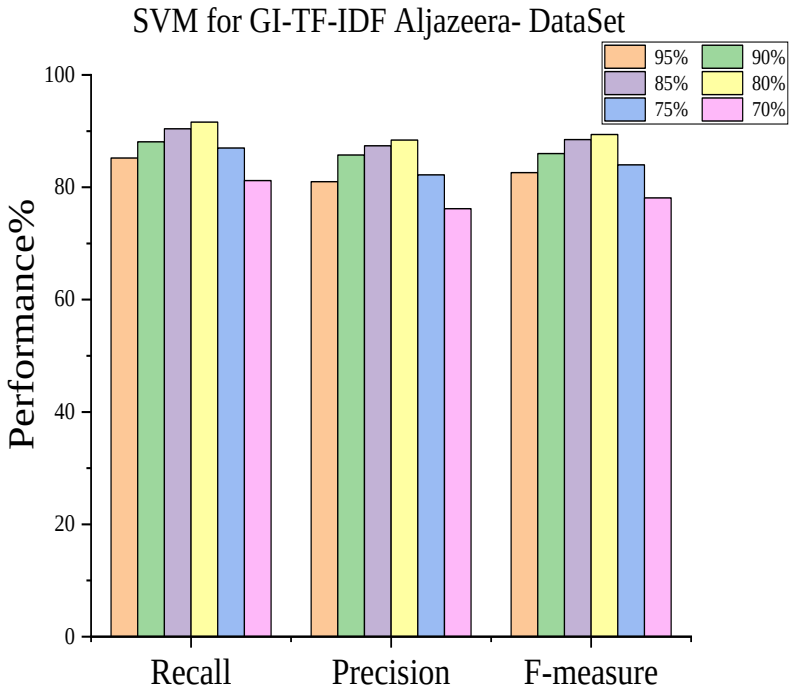




❑ Experimental results (cont.)

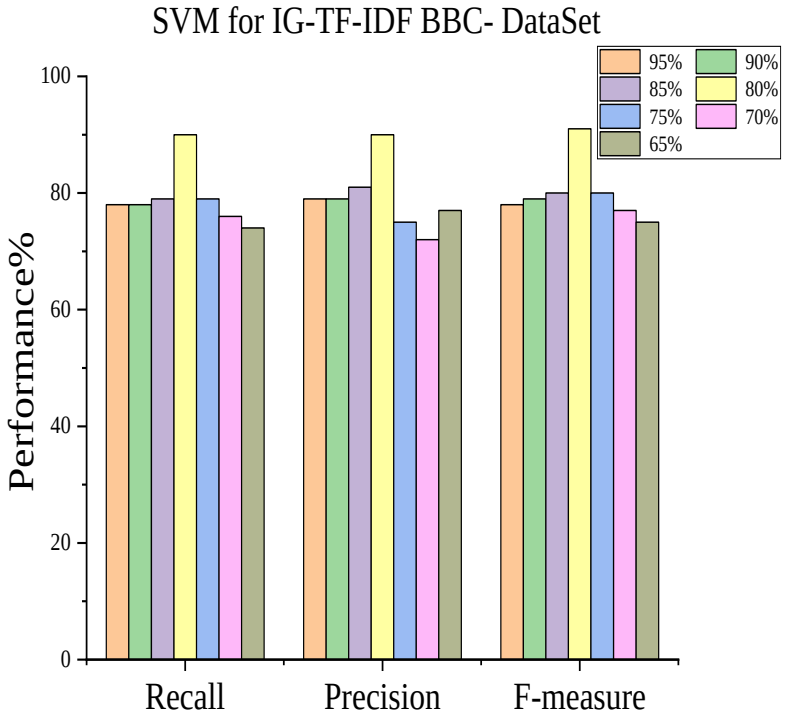
➢ Enhancement Using Combination of Gini Index and TF-IDF Feature Selection

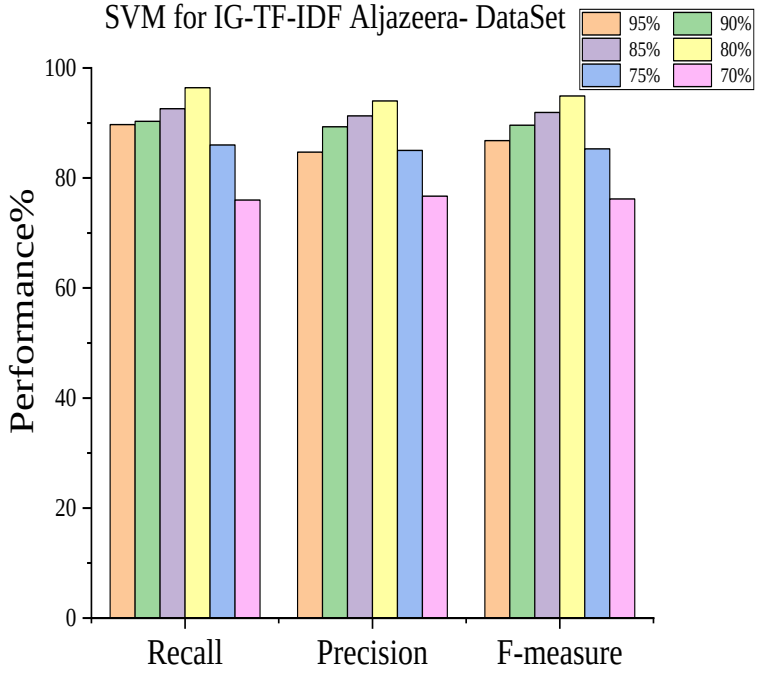




❑ Experimental results (cont.)

➢ Enhancement Using Combination of Information Gain and TF-IDF Feature Selection





❑ Proposal Feature Selection Method

➢ Semantic Fusion Method

- We intend to propose a feature selection approach based on using two axioms namely: the multiple words and the word senses.

- Amalgamating two or three sequence of words in a document instead of using such individual words will reduce the number of features. For example, the three individual words 'جمهورية', 'مصر', 'العربية' can be concatenated together to form one feature in a phrase form or a multiple-words form.

❑ Proposal Feature Selection Method (cont.)

➢ Semantic Fusion Method (cont.)

- The collection of documents is represented in a matrix with size N×S where N is the number of instances while S is the number of features describing the documents.

- If two or three individual terms are used as features and those individual terms are included in a long multiple word feature, the individual features are eliminated from the feature set.

❑ Proposal Feature Selection Method (cont.)

➢ Semantic Fusion Method (cont.)

- After computing the similarity among the individual features the result will be put in a matrix of size N×S.

	f_1	f_2	f_3	…..	f_n
d_1	w_{11}	w_{12}	w_{13}		w_{1S}
d_2	w_{21}	w_{22}	w_{23}		w_{2S}
d_3	w_{31}	w_{32}	w_{33}		w_{3S}
. .					
d_N	w_{N1}	w_{N2}	w_{N3}		w_{NS}

❑ Proposal Feature Selection Method (cont.)

➢ Semantic Fusion Method (cont.)

- The highly related features can determine the same category or class; so one of such features can be eliminated while keeping the other one as a feature in the feature set.
- The S features describing the dataset can be considered as S vectors. The similarity measure $Sim(f_i,f_j)$ between any two features f_i and f_j can be computed as shown in equation.

$$Sim(f_i, f_j) = \frac{\sum_{i,j=1}^{N} f_i f_j}{\sqrt{\sum_{i=1}^{N} f_i^2}\sqrt{\sum_{j=1}^{N} f_j^2}}$$

As a result, the matrix represents the similarity values among the individual features as shown.

❑ Proposal Feature Selection Method (cont.)

➢ Semantic Fusion Method (cont.)

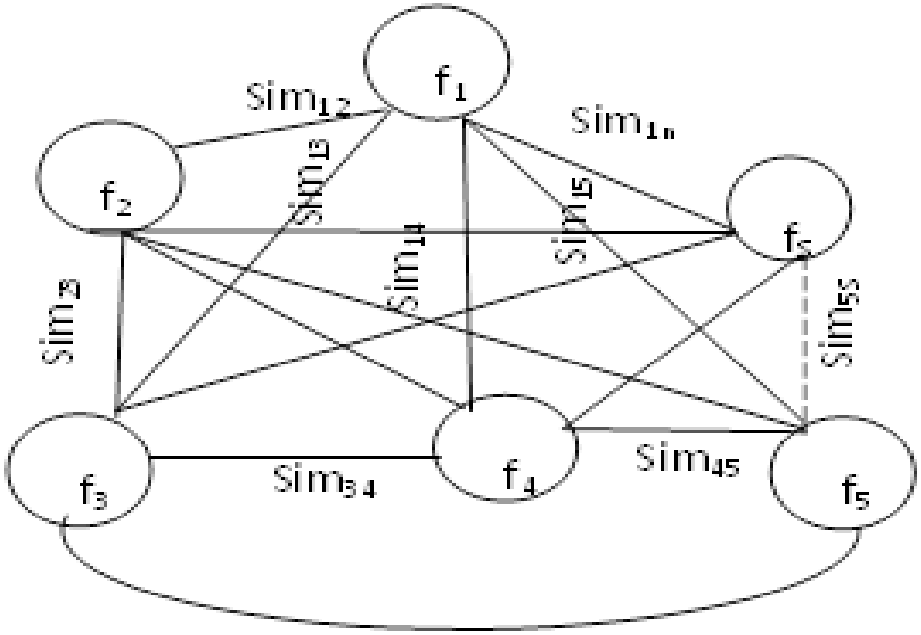
- As a result, the matrix represents the similarity values among the individual features as shown in this figure.

	f_1	f_2	f3	……f_S
f_1	1	Sim_{12}	Sim_{13} . . … ...Sim_{1S}	
f_2		1	Sim_{23} …….. .Sim_{2S}	
f_3			1 …….. Sim_{3S}	
⋮				
				f_S

❑ Proposal Feature Selection Method (cont.)

➢ Semantic Fusion Method (cont.)

- The similarity values among the features can be represented in a graph consisting of S nodes connected together by links as shown



❑ Proposal Feature Selection Method (cont.)

➢ Semantic Fusion Method (cont.)

- The number of features can be reduced by adopting the fusion concept as in the following steeps.
- Let us assume a starting threshold value T_s.
- If the similarity value Sim_{12}, or Sim_{13}, ….., or Sim_{1S} is less than the threshold value T_s then those corresponding features are taken into consideration.
- All the connected features with a certain feature say f_1 with similarity values greater than or equal T_s can be eliminated.

❑ Proposal Feature Selection Method (cont.)

➢ Semantic Fusion Method (cont.)

- In this case f_1 is considered a combined feature and it is added to the features set.
- The number of features becomes the original S features without those eliminated ones.
- This process is repeated till the last feature f_S.
- The resulted number of features will be the input to the classifier.
- Other experiments are run and operated for other threshold values.

❑ Proposal Feature Selection Method (cont.)

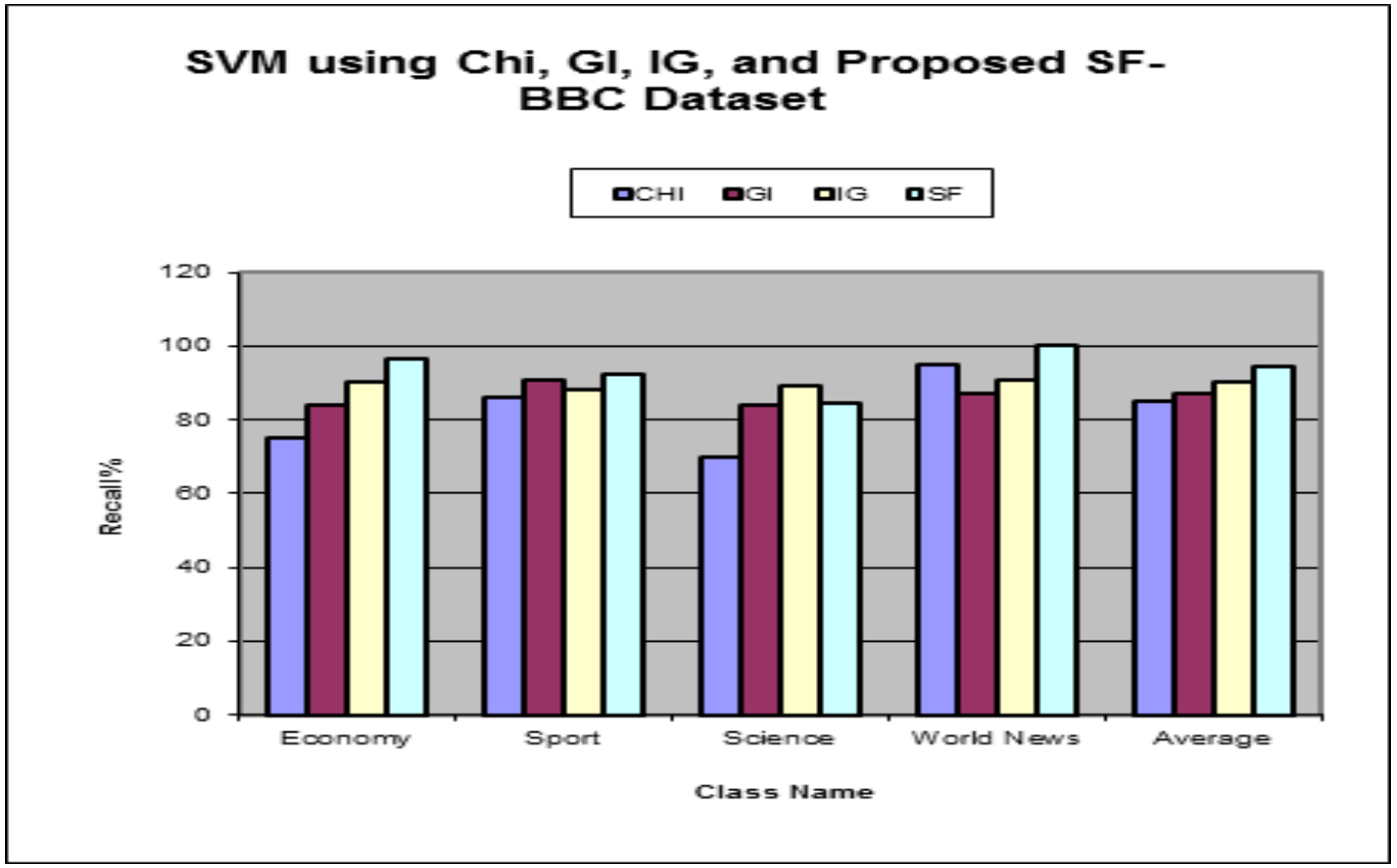
➢ Semantic Fusion Method (cont.)

Algorithm: Combined Semantic Features Fusion
Input: Matrix (N,S) / * N= the no. of documents and S is the no. of features*/
Output: No. of features after combining or fusing the features
Steps:

Compute the similarity values among the individual features
Set a starting threshold T_s to a specifie value
Set K=0 /* K is the number of eliminated features*/
Set R=0
While (T_s < Sim_{max})
For i=1 to N
For j=i+1 to (N-R) Do
If Sim (i,j) < T_s
Then K=K+1
Next j
R=K
K=0
Next i
End While
The no. of features after fusion= S-R

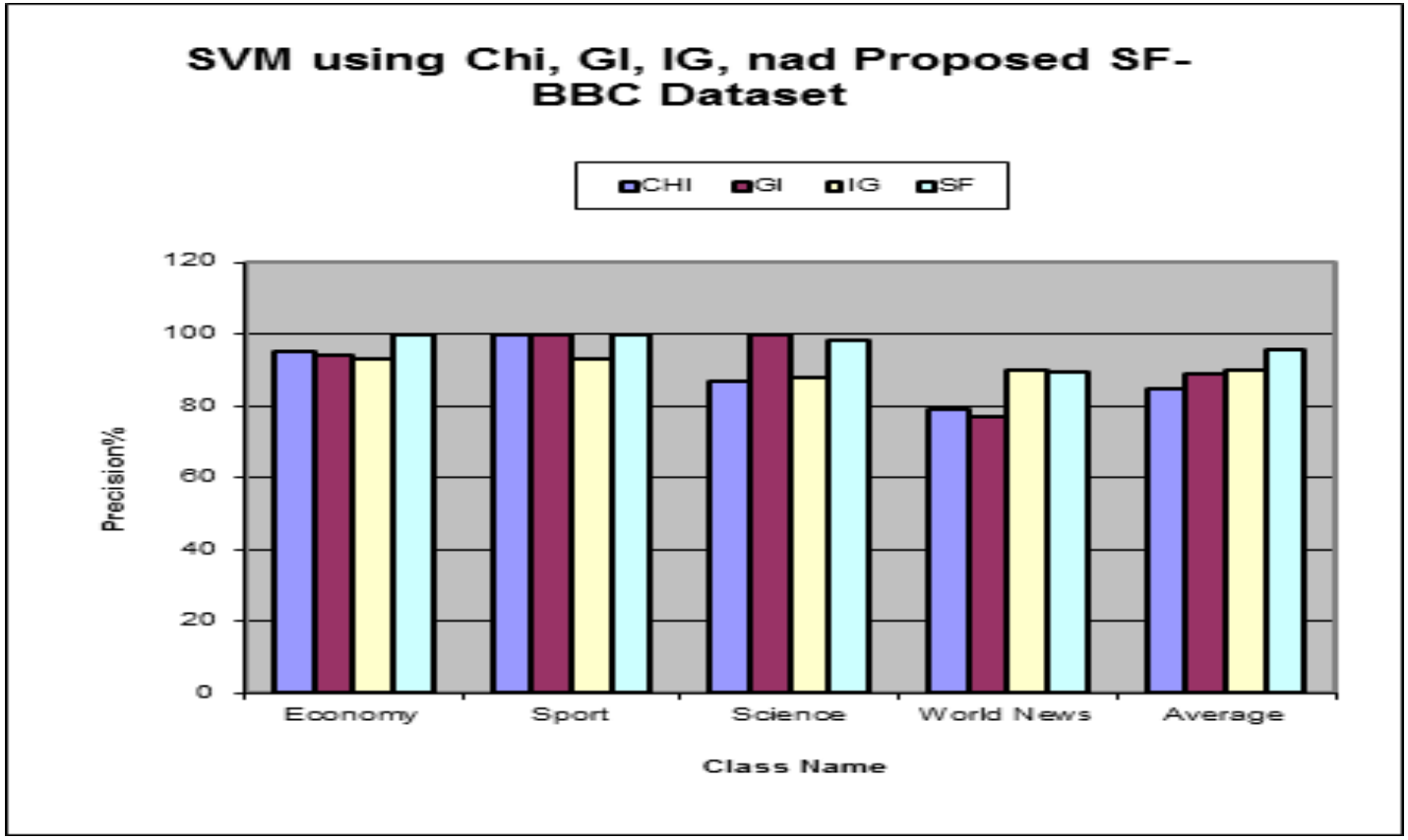
❑ Experimental results (cont.)

- Recall% using All methods on BBC Dataset



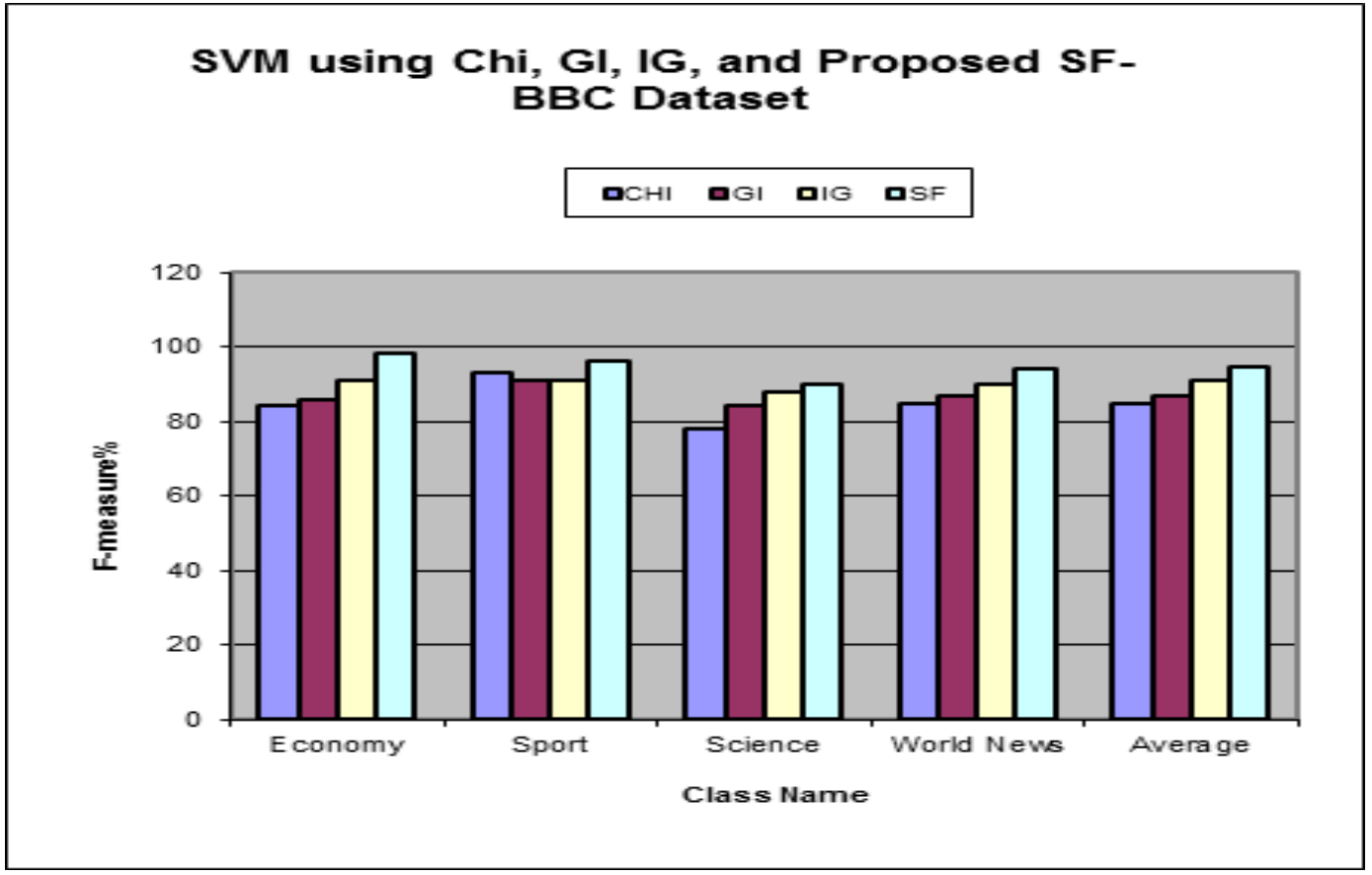
❑ Experimental results (cont.)

- Precision % using All methods on BBC Dataset



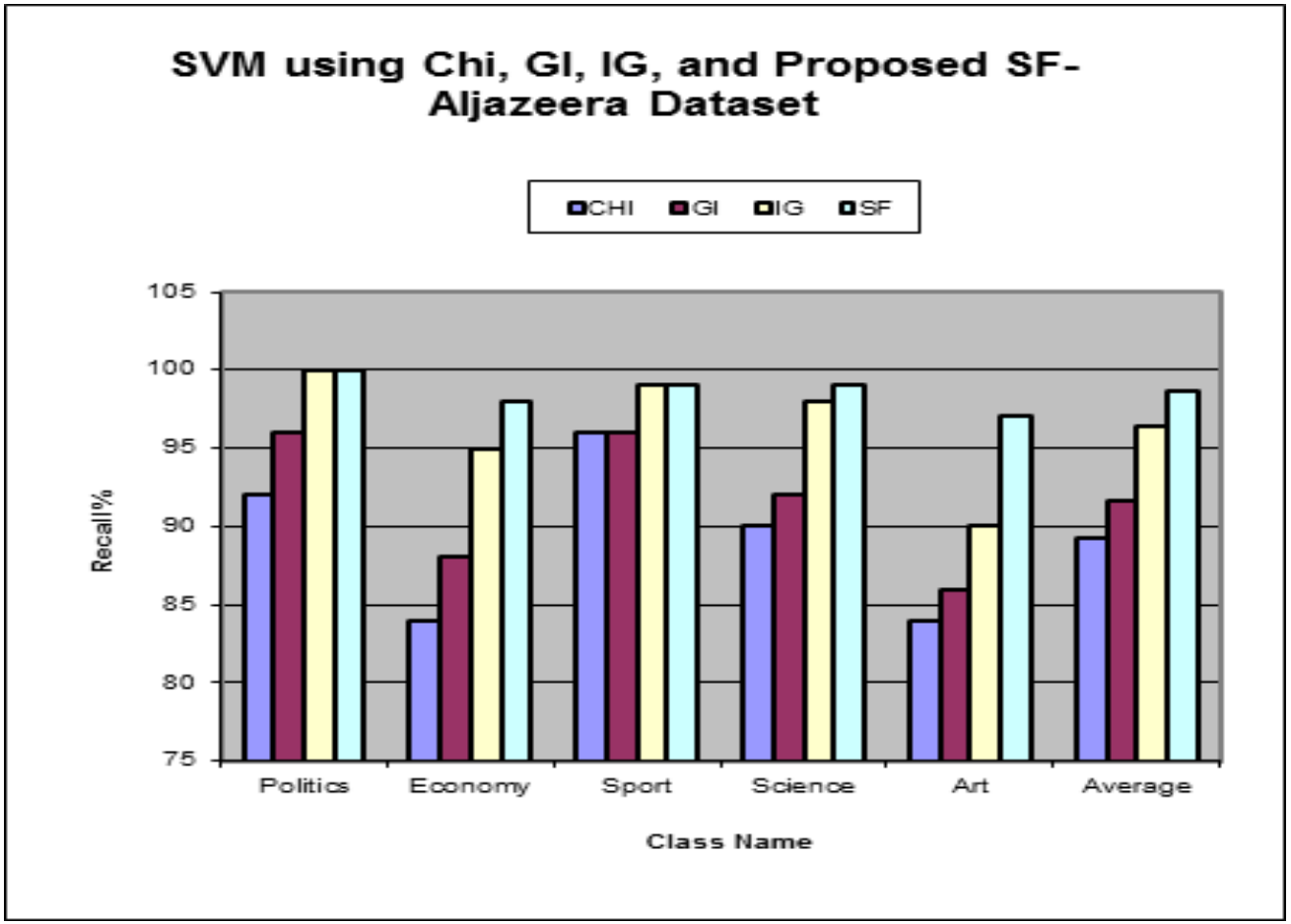
❑ Experimental results (cont.)

- F-measure% using All methods on BBC Dataset



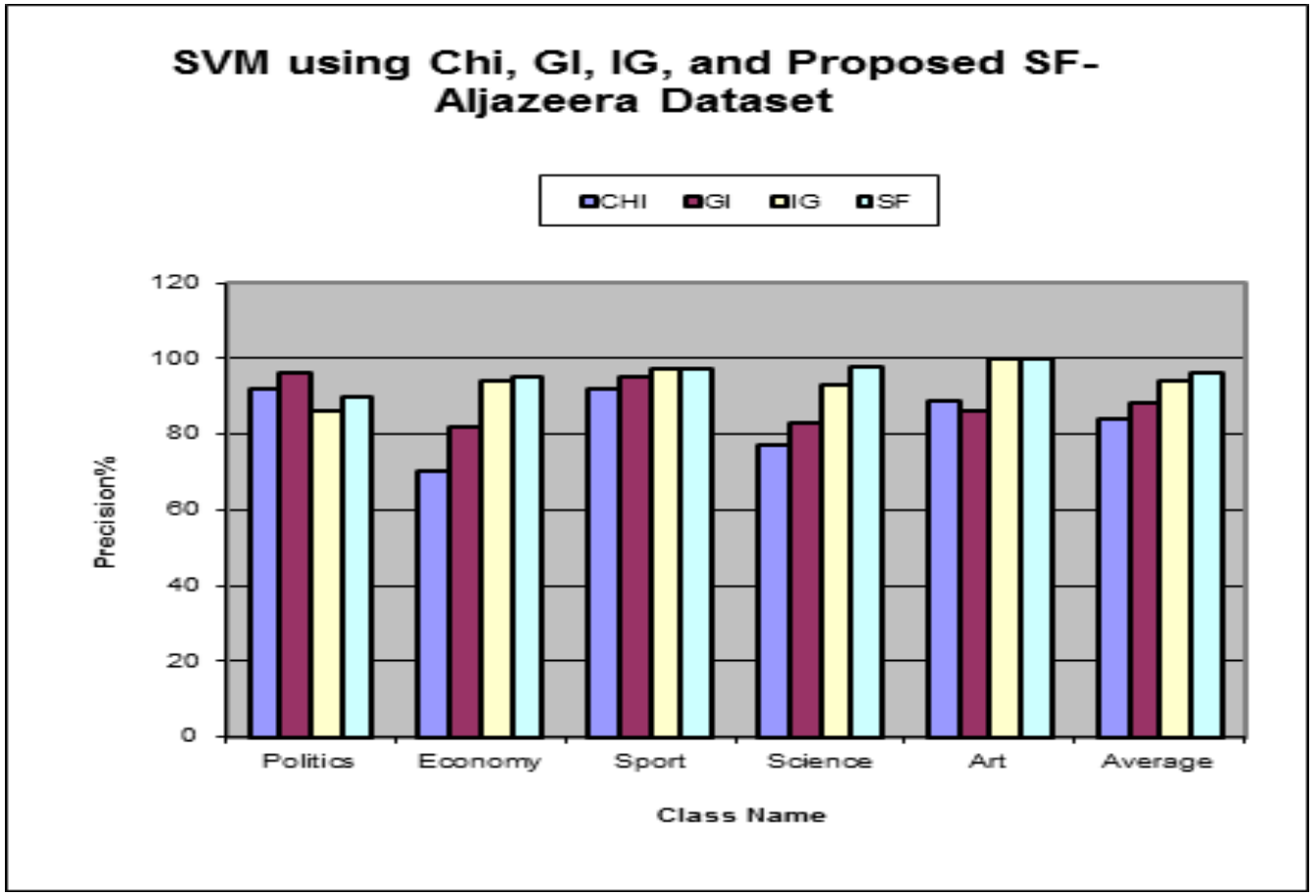
Experimental results (cont.)

- Recall% using All methods on Aljazeera Dataset



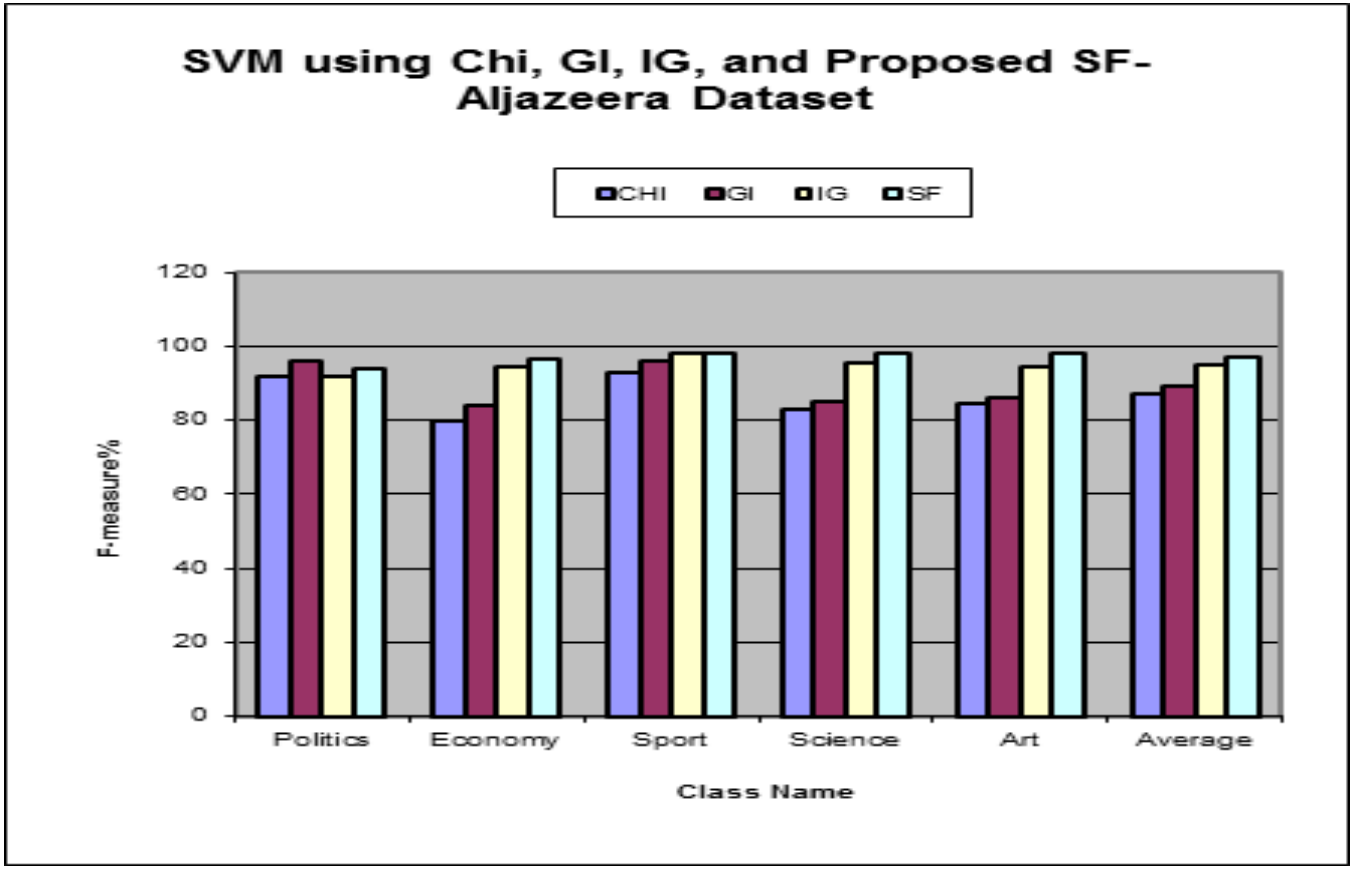
❑ Experimental results (cont.)

- Precision% using All methods on Aljazeera Dataset



❑ Experimental results (cont.)

- F-measure% using All methods on Aljazeera Dataset



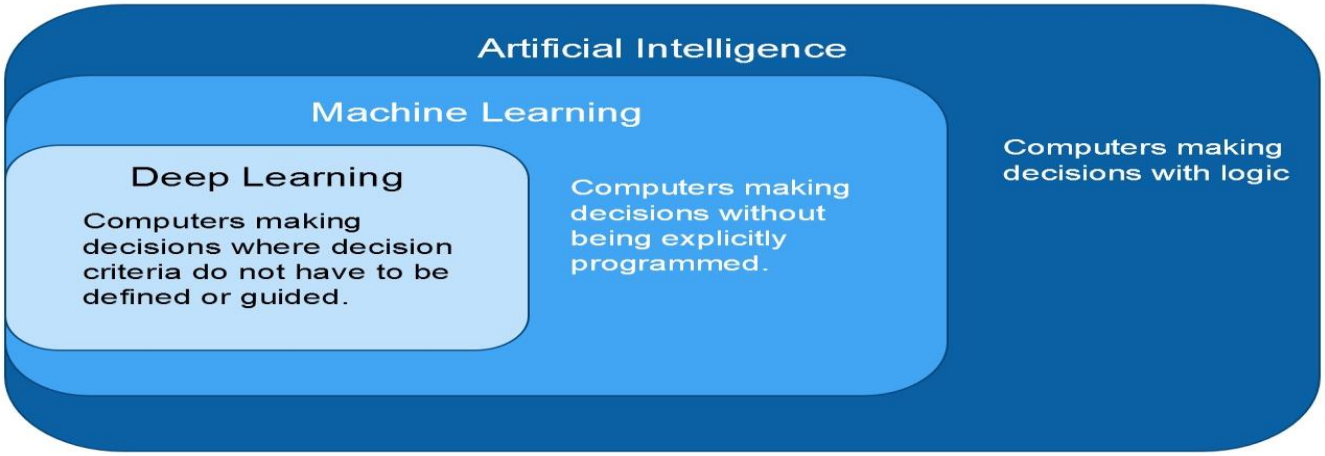
❑ Experimental results (cont.)

	BBC Arabic News Dataset			Aljazeera Arabic Dataset		
Previous Literature	Recall%	Precision%	F-Measure%	Recall%	Precision%	**F-Measure %**
[Ibrahim Abuhaiba and Hassan Dawoud, 2017][6]	88.2	87.9	88.1	--	--	--
[Ahmed T. Abdulameer et.al, 2017] [2]	72.0	71.0	71.0	--	--	--
[Hamza Mohammed Naji, 2016] [4]	82.9	83.4	85.2	89.7	84.7	86.8
[W.A. Awad, 2012] [14]	--	--	--	86.1	78.1	81.9
[Adel Hamdan, et.al, 2016] [1]	--	--	--	77.4	77.8	77.5
[Mayy M. Al-Tahrawi, 2016] [9]	--	--	--	84.0	85.0	84.3
Combining chi+TF-IDF	85.0	85.0	85.0	89.2	84.0	**87.0**
Combining GI + TF-IDF	87.0	89.0	87.0	91.6	88.4	**89.4**
Combining IG + TF-IDF	**90.0**	**90.0**	**90.0**	**96.4**	**94.0**	**94.9**
Proposed SF-MW	**94.5**	**95.4**	**94.5**	**98.6**	**96.0**	**96.9**

❑ A Chosen Deep Learning Model

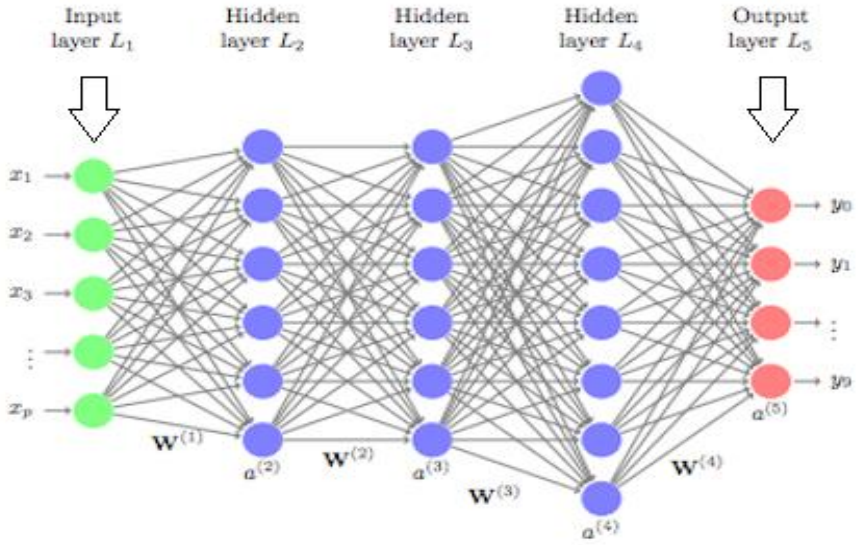
- Deep learning is a class of machine learning algorithms that use a cascade of layers of processing units to extract feature from data.

- The concept of deep learning was originated from the study of artificial neural networks which are inspired by biological brain model made of neurons.



❑ A Chosen Deep Learning Model(cont.)

- The deep learning used to enhance and improve the results of the traditional types of machine learning.
- One of the most popular types of deep neural networks is known as Convolution Neural Network (CNN)



❑ Convolution Neural Networks

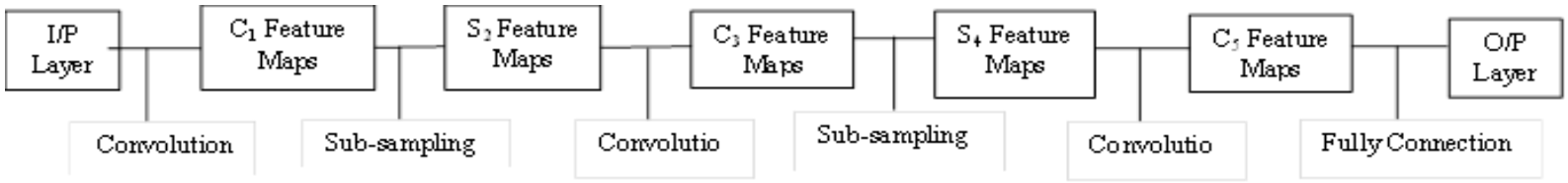
- The function of CNN is learning special features of the data and then makes convolution down to a smaller subset of the data.
- The CNN can perform a mathematical transformation known as convolution to the input data.
- The convolution can be achieved by passing a filter over the input matrix representation of the document.
- Filters are a matrix of parameters or weights where the numerical values are determined in the process of model fitting.

❑ Convolution Neural Networks (cont.)

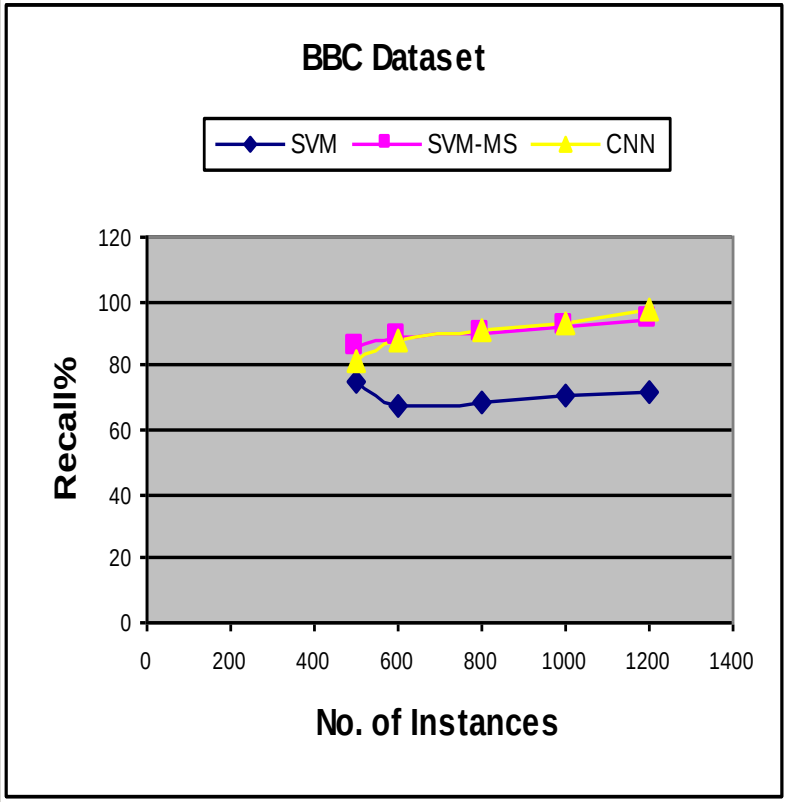
- The CNN as a multi-layer neural network contains four type of layers:

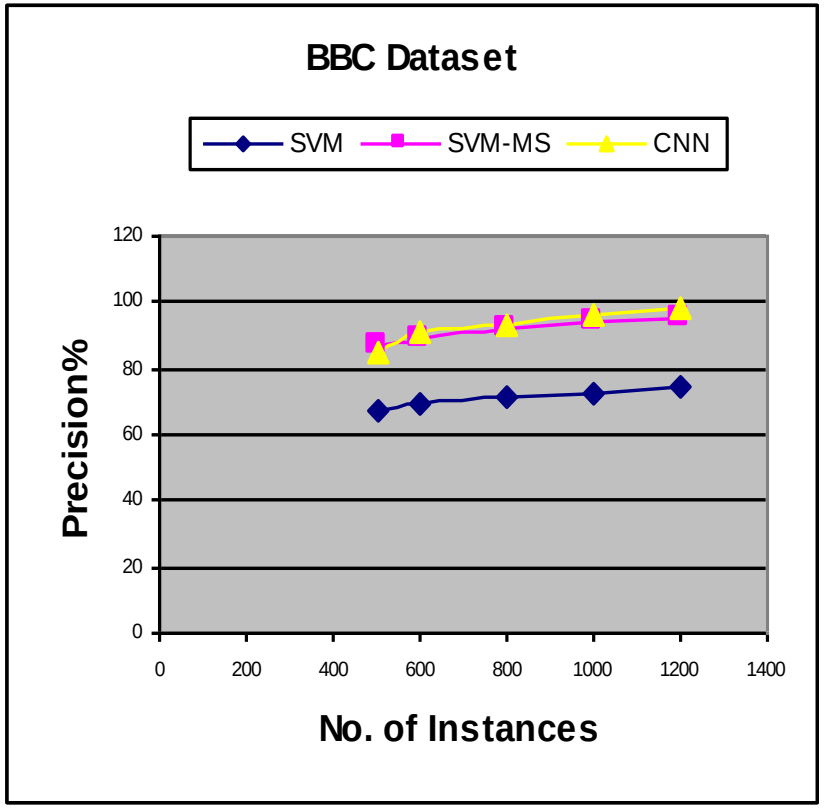
- convolution layers (C-layers) can be used to extract features when the input of each neuron is linked to the local receptive field of the previous layer.
- Max pooling (Sub-sampling)(S-layers)

  It uses Max function to reduce the number of Features.
- Flattening
- Full Connection

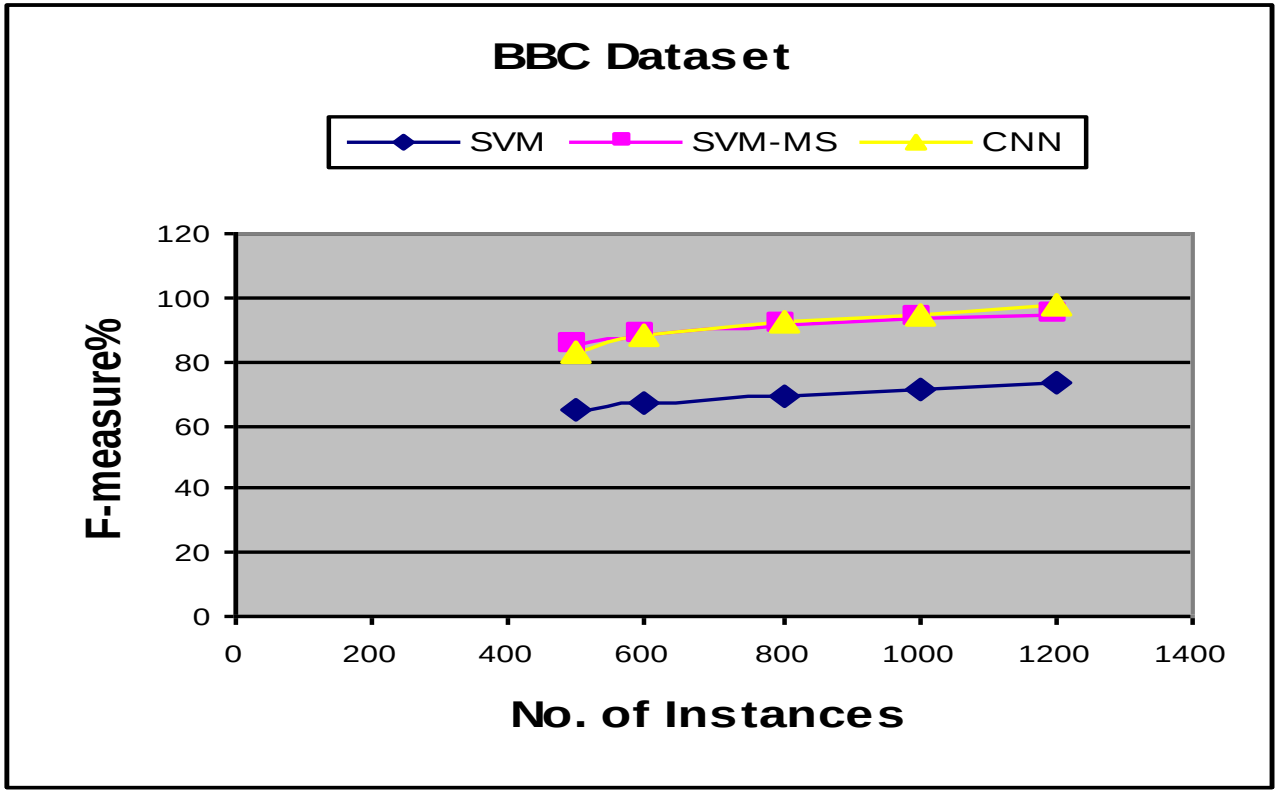


Experimental Results



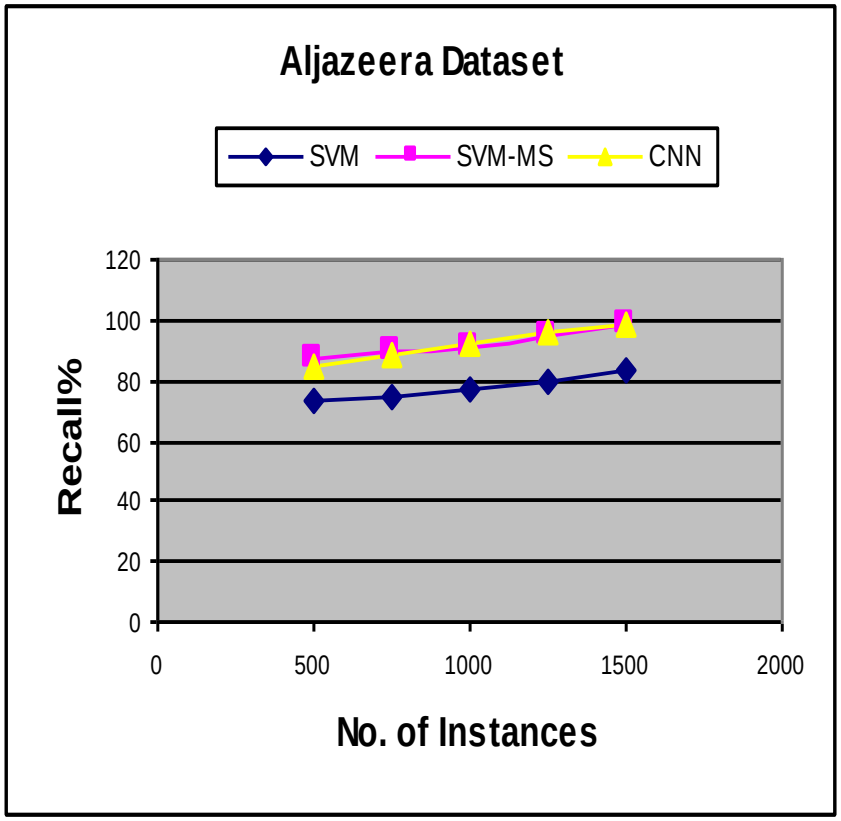


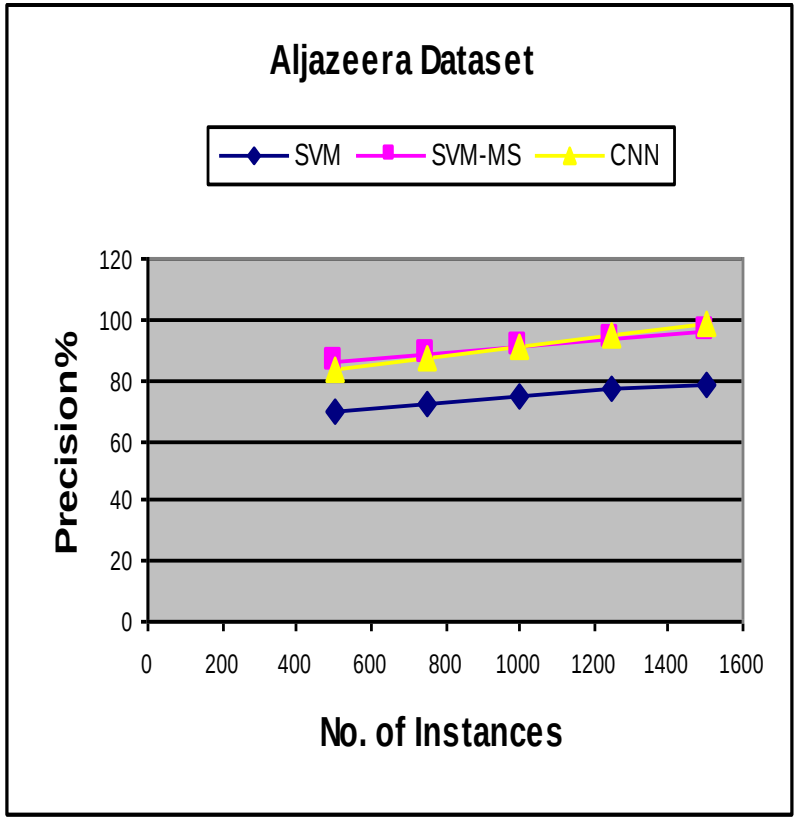
❑ Experimental Results (cont.)



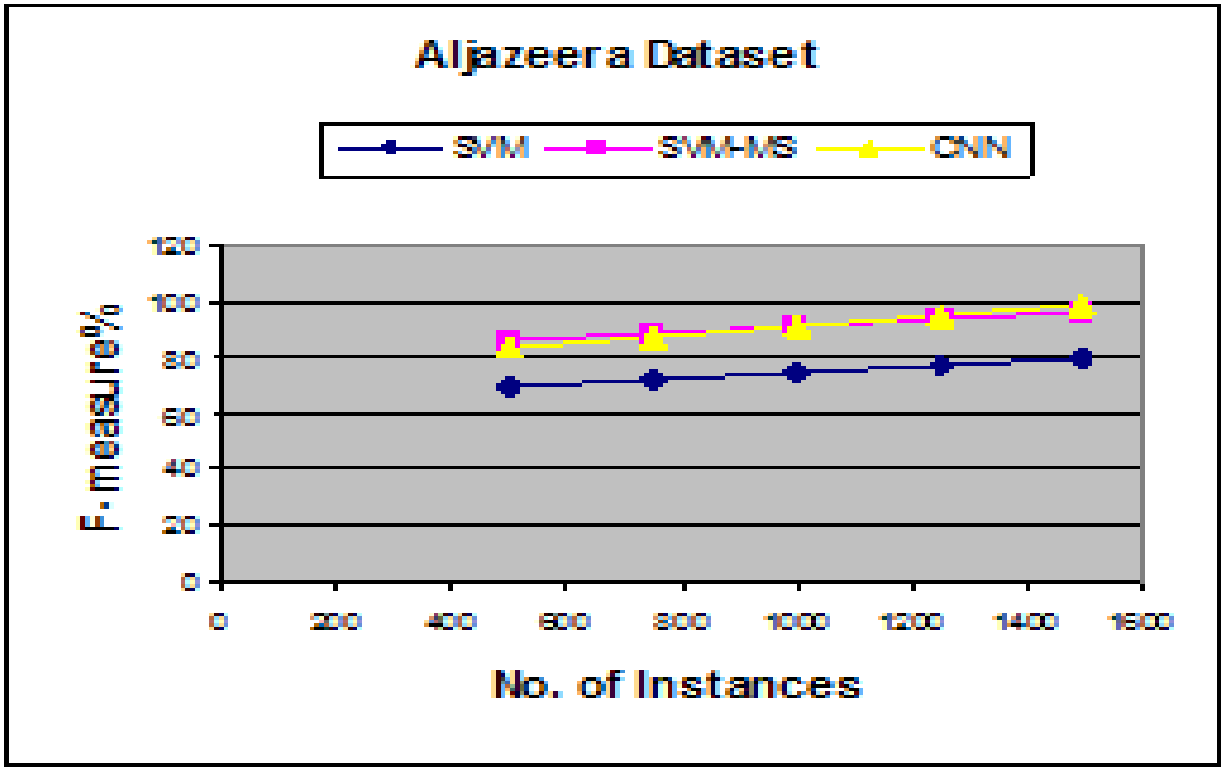
Approach	Improve
SVM-MS	21%
CNN	25%

❑ Experimental Results (cont.)



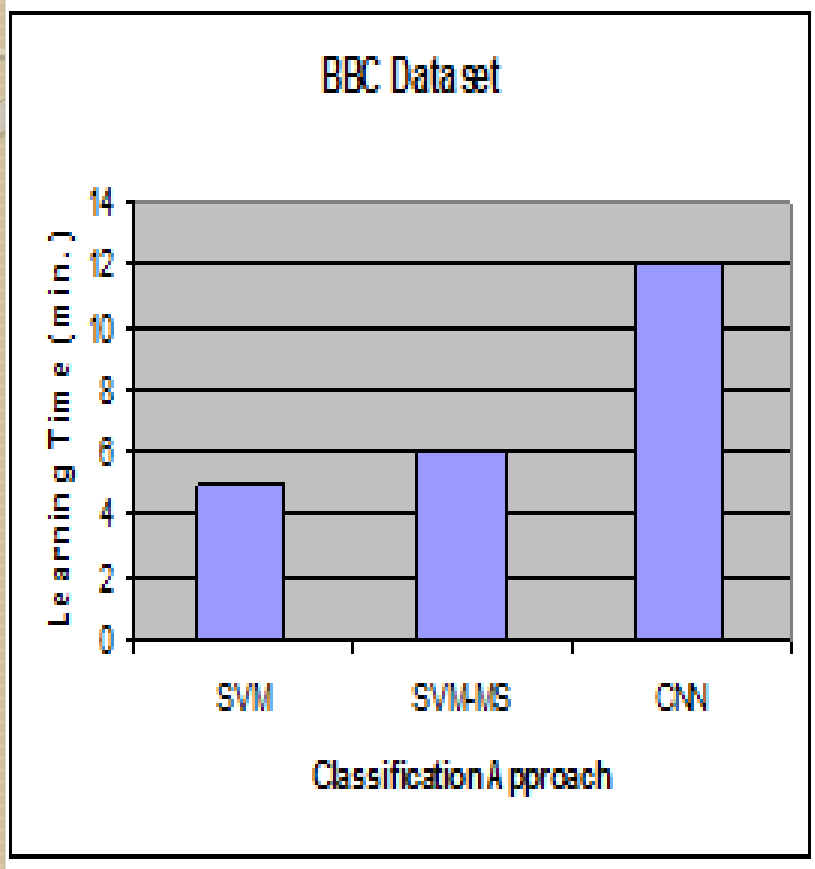


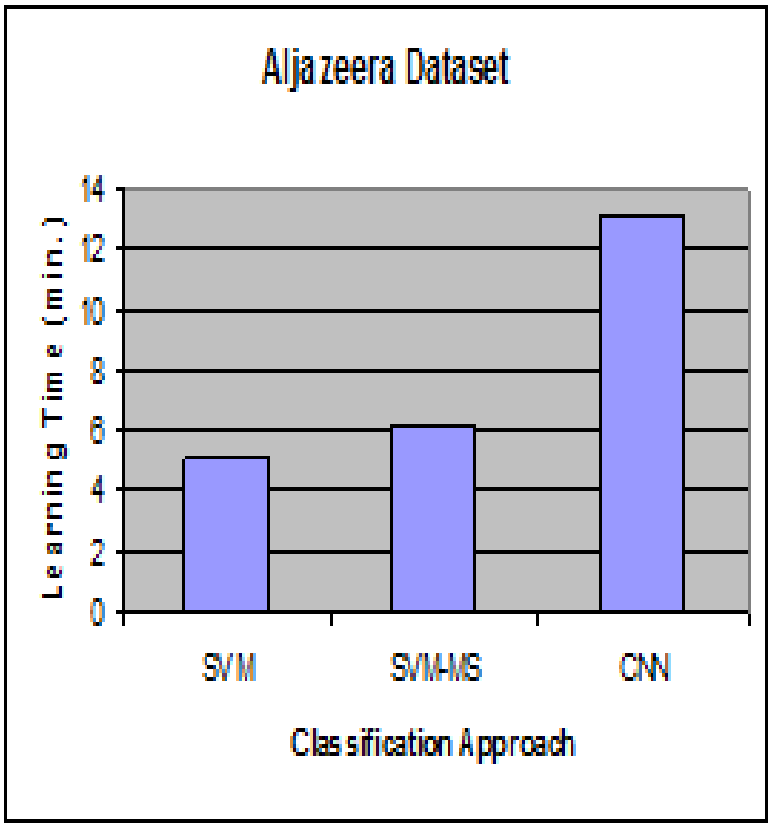
❑ Experimental Results (cont.)



Approach	Improve
SVM-MS	16%
CNN	18.5%

❑ Experimental Results (cont.)





❑ Conclusion and Future Work

➢ Conclusions

- Due to the increasing widespread of the World Wide Web; the area of information retrieval systems is important.
- Information retrieval involves representation, storage, and accessing of information which are stored on document collections.
- Information retrieval systems aim to retrieve the relevant documents of unstructured data that satisfy the user needs.

❑ Conclusion and Future Work (cont.)

➢ Conclusions (cont.)

- An Arabic information retrieval model was developed, implemented, and evaluated. The model has several modules namely: Arabic collection of documents, query processing, document-query matching, indexing, and ranking the retrieved documents.
- The performance of the Arabic model was tested using some measurable evaluation criteria such as precision, recall, and F-measure.
- The query processing operation was modified by reformulating the user queries using the semantics of keywords and using also the relevance feedback.

❑ Conclusion and Future Work (cont.)

➢ Conclusions (cont.)

- Due to the query reformulation; the number of relevant retrieved documents was increased. The values of precision, recall, and F-measure were better than those corresponding ones without query reformulation.
- The query reformulation using the semantics of query keywords and the relevance feedback approaches improved the performance by about 27% and 14% respectively. Combing the two approaches together; the performance outperforms the retrieval model without modification by about 15% to 35%.

❑ Conclusion and Future Work (cont.)

➢ Conclusions (cont.)

- Query expansion was done using the word embedding concept which is based on word2vec model.
- Word2Vec model represents each word W of the training set as a vector of features and that vector can capture the context in which W appears.
- Two models; continuous-bag-of-words (CBOW) and Skip-Gram; and the hybrid of the both were discussed, operated and evaluated for expanding the users' queries. The models were tested using the standard CNN dataset which contains about 92000 documents.

❑ Conclusion and Future Work (cont.)

➢ Conclusions (cont.)

- The vector length, window size, the number of expanded terms for each query keyword, and the expansion approaches have significant effects on the overall performance.

- The best performance occurred for the vector length equals three-hundreds, and the window size equals five. The best performance also occurred when expanding each query keyword by only one candidate term; and sometimes two terms.

❑ Conclusion and Future Work (cont.)

➢ Conclusions (cont.)

- The performance of CBOW model was better than that of the Skip-Gram. The hybrid of CBOW and Skip-Gram presented the best performance because it utilized the benefits of both.
- The number of good matched expanded terms for the hybrid approach was greater than the corresponding number of either the CBOW or the Skip-Gram.
- The performance of query expansion using word embedding was better than that performance without expansion. The improvement in performance was achieved by about 24%, to 36% for the CBOW compared with some of those in literature. Also, the performance of CBOW outperforms the Skip-Gram by about 31%. The hybrid approach outperforms the performance of CBOW by about 17.14%.

❑ Conclusion and Future Work (cont.)

➢ Conclusions (cont.)

- Text or document classification is important for a lot of applications.
- Three classification approaches were analyzed, operated, and evaluated. The approaches are decision tree (DT), Naïve Bayes (NB), and support vector machine (SVM).
- The performance of the classifiers was tested using two standard datasets namely: BBC Arabic and Aljazeera datasets. The SVM classifier presented the best results compared with the performance of the other two classifiers.
- Several feature selection methods were analyzed, discussed, and evaluated. The methods are based on: term weighting, information gain, Gini index, and chi-square.

❑ Conclusion and Future Work (cont.)

➢ Conclusions (cont.)

- The classification accuracy was improved due to the change in the number of selected features. The performance of the amalgamated feature selection methods was better than that performance of every method individually.
- A feature selection method was proposed. The method is based on semantic feature fusion and multiword (abbreviated as SF-MW). Some features were fused and multi-words were considered as one feature instead of using the individual words as many features. The fusion reduced the number of selected features.
- The performance of the proposed method outperforms the other adopted ones. The improvement of the performance was up to 22% and that improvement occurred for the two chosen Arabic datasets.

❑ Conclusion and Future Work (cont.)

➢ Conclusions (cont.)

- A deep learning approach based on convolution neural network (CNN) was adopted and discussed.
- A comparative study was done for three approaches for text classification. The approaches are the original SVM, the modified SVM based on SF-MW and the CNN. The approaches were tested using two test-bed datasets.
- The performance of the CNN deep learning was the best while the worst was the SVM. The performance of the SVM based on SF-MW was better than SVM by about 23%, 21%, and 21% respectively for precision, recall, and F-measure using the BBC dataset. Also, its performance was better than the SVM by about 15.5%, 18%, and 17% respectively for the some metrics using Aljazeera dataset.

❑ Conclusion and Future Work (cont.)

➢ Conclusions (cont.)

- The performance of CNN deep learning approach was effective and robust due to the good results for using multiple convolution layers.
- If the number of layers increased above a threshold number, the performance metrics begin to decrease. That degradation may be occurred due to the information loss during learning operation.
- The performance of CNN deep learning approach was improved by about 25%, 24%, and 24% respectively for recall, precision, and F-measure compared with the SVM classifier.

❑ Conclusion and Future Work (cont.)

➢ Conclusions (cont.)

- It is expected that the deep learning approach may give more improvement in the performance metrics for bigger sizes of datasets.

- The learning time was the largest for deep learning and smallest for the SVM classifier.

❑ Conclusion and Future Work (cont.)

➢ Future Work

The following titles are considered as key points for future work directions.

- Text processing and mining for social media monitoring
- Arabic intelligent web search engine
- Document analysis using new trends of big data analytics
- Text processing using quantum computing approaches
- Intelligent text summarization
- Text to speech conversion-speaker dependent

❑ Liste of Publicaciones

[1] Ayat Elnahas, Nawal A. El-Fishawy, Mohamed Nour, Gamal Attiya and Maha Tolba,"Query Expansion for Arabic Information Retrieval Model: Performance Analysis and Modification", Egyptian Journal of Language Engineering, Vol. 5, No. 1, PP. 11-24, April 2018.

[2] Ayat Elnahas, Mohamed Nour, Nawal A. El-Fishawy and Maha Tolba, "Machine Learning and Feature Selection Approaches for Categorizing Arabic Text: Analysis, Comparison, and Proposal", Egyptian Journal of Language Engineering, Vol. 7, No. 2, PP. 1-19, 2020.

[3] Ayat Elnahas, Mohamed Nour, Nawal A. El-Fishawy and Maha Tolba, "Performance Evaluation of Some Query Expansion Methods Using Word Embedding", to Appear.


Thank yow