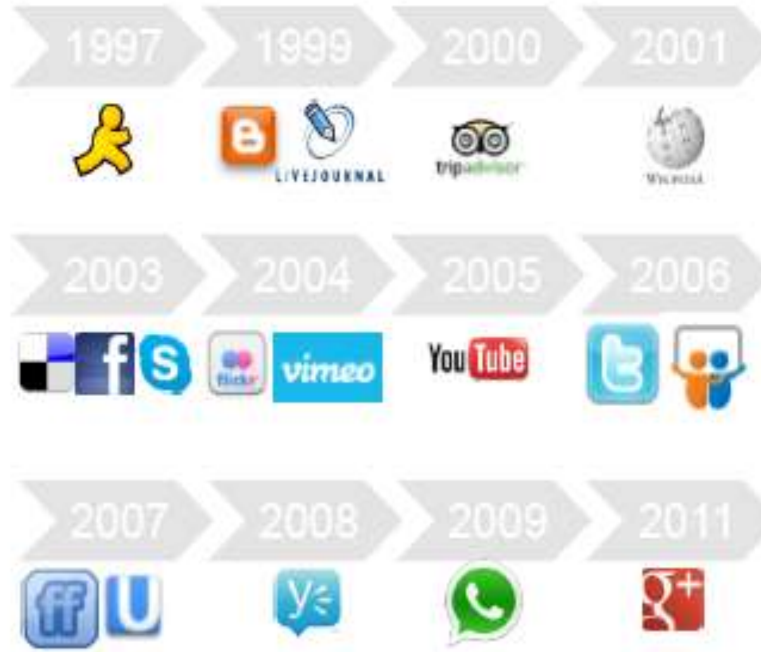
Séminaire de l'équipe Linguistique Computationnelle

Ibtihel BEN LTAIFA

A New Representation and Ranking Approaches based on Deep Learning to Improve the Semantic Information Retrieval in Microblogs

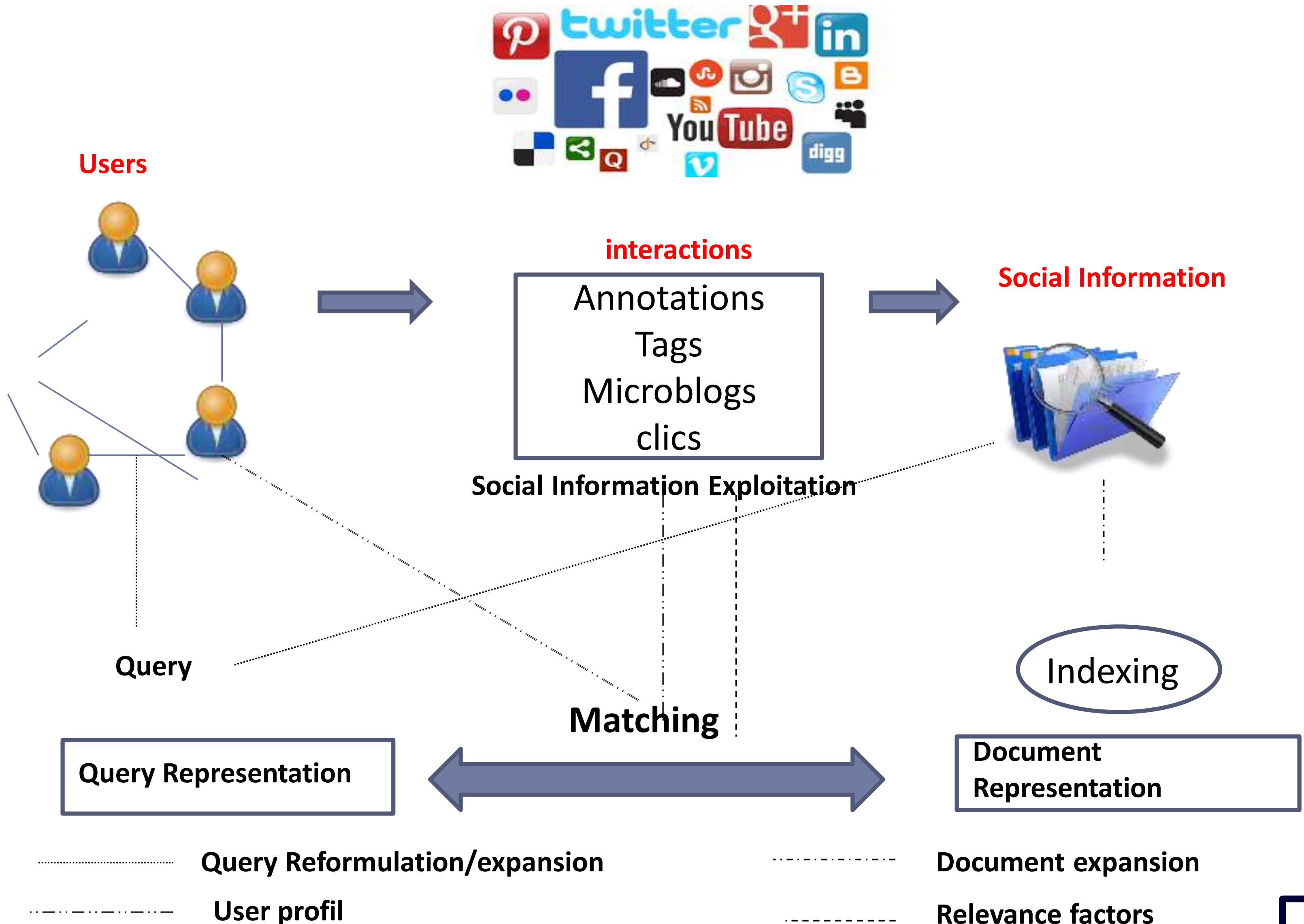
Plan

Web 2.0

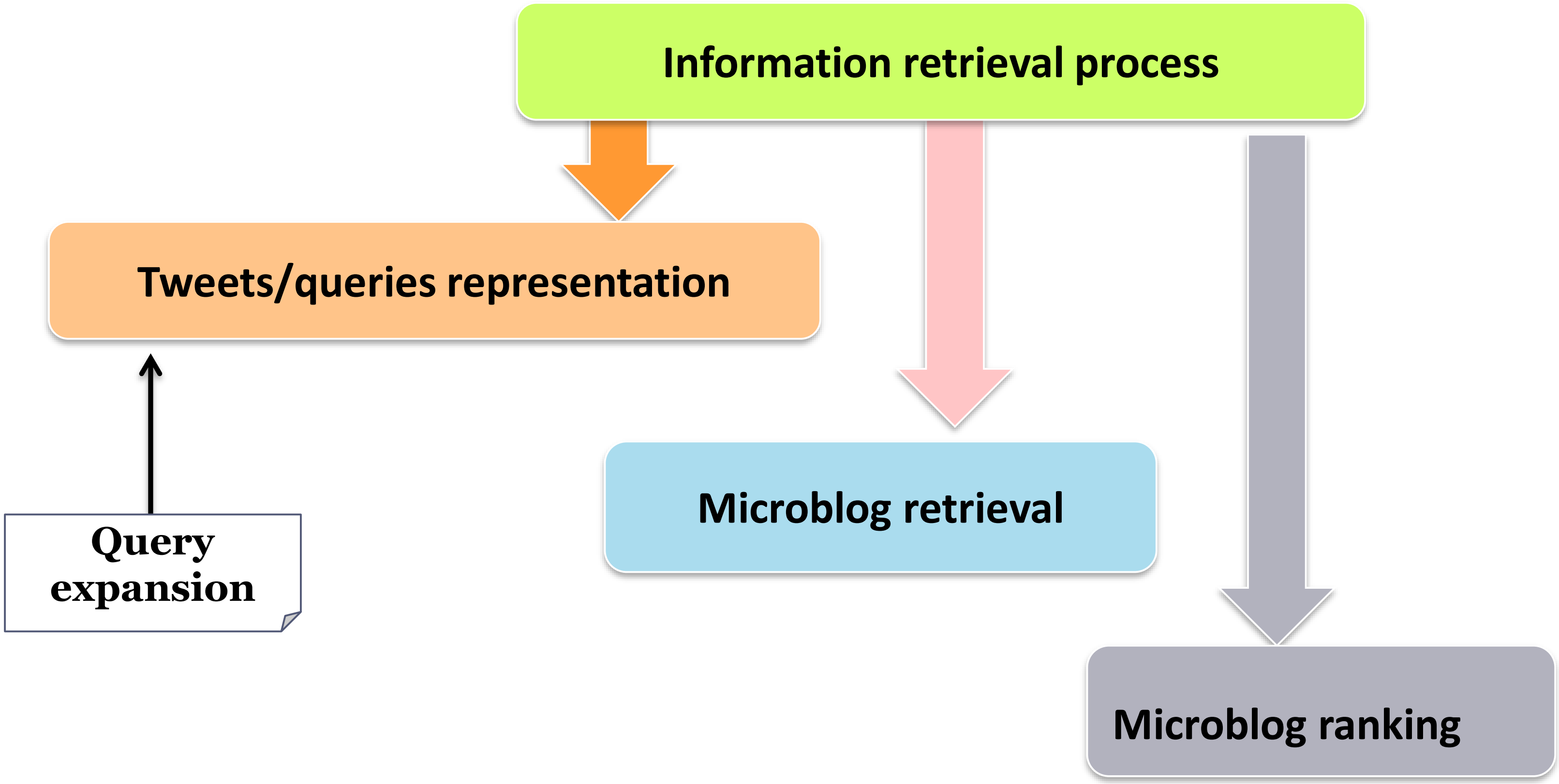


-communicating
-interacting
-publishing news
-sharing ressouces
-exchange messages
-commenting statuses
-creating profiles through these platforms

Social Information Retrieval: context



Social Information Retrieval: Challenges



Microblogs specificities :

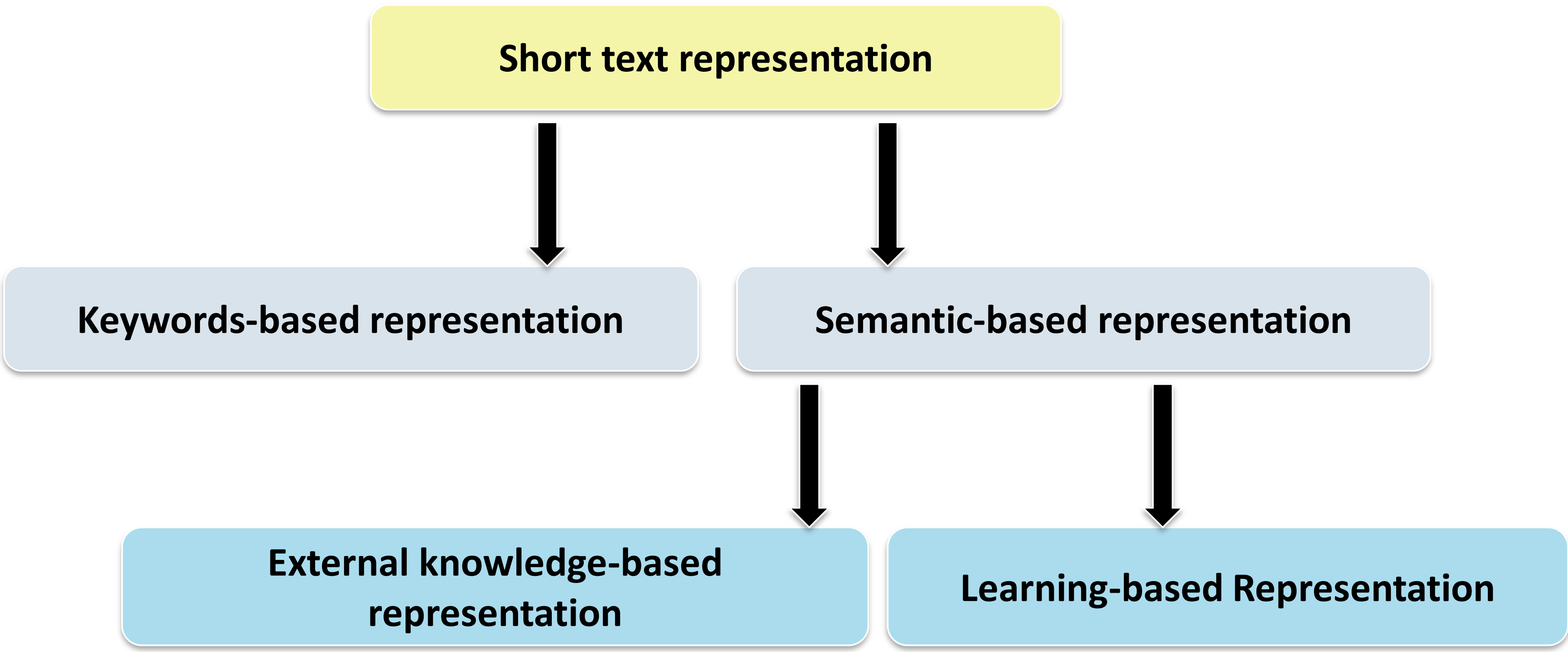
-Short messages with limited characters (up to 280 characters in Twitter)

-Specific syntax (@mention, #hashtag, RT…)

-May contain URLs

-Quality of language (poor syntax: misspelled terms, abbreviations,…)

All of these specificities of microblogs introduce new challenges !


Short text representation
Keywords-based representation
Semantic-based representation
External knowledge-based representation
Learning-based Representation

Keywords-based representation

•**BoW** [Bansal& al., 2015] [Ferguson & al., 2012] [Lin & al., 2012]

-High dimensional feature vector

-Term ordering is not considered

-Cannot capture semantics

- Ignores relationships between words .

Polysemy and lexical ambiguity problems!

Semantic-based representation

External knowledge-based representation

- **Knowledge Base (DBpedia)**

[Fahd &al., 2016] [zingla &al., 2018]

-Focus on concept similarity rather then senses

-Penalize synonymous relationships that co-occur for the same object

Learning-based Representation

- **Word embeddings (word2vec)**

[Moumit&al.,2017] [Roy &al.,2016]

-Do not solve the problem of polysemous words

-The sub-linear relationships are not defined

-Large features dimension

- **Microblog retrieval (Matching)**

Classical models (BM25, Boolean)

- ❑ Based on factors such as the frequency of terms in documents: **ineffective** with short text [Choi &al, 2012] [Damak&al., 2014]
- ❑Based on exact term matching
- ❑Originally designed for long text
- ❑ No longer adapted to the specificities of the new form of content in microblogs

Vocabulary mismatch [Damak&al.,2013]

- ❑ Conciseness of microblogs
- ❑Total absence of the terms of queries
- ❑Named entities recognition
- ❑Abbreviations written in different ways

Relavance

Textual Relevance

- Only textual content [Che Alhadi&al., 2011]

Social Relevance

•Incorporating social factors in the relevance score :
[Patel&al.2017] [Ben Jabeur&al., 2012]

-freshness

-number of followers

-popularity of the author

-number of retweets

Which social factor is more effective to improve relevance?

Social Information Retrieval: Challenges

Relavance

Textual Relevance

- Only textual content [Che Alhadi&al., 2011]

Social Relevance

•Incorporating social factors in the relevance score :
[Patel&al.2017] [Ben Jabeur&al., 2012]

-freshness

-number of followers

-popularity of the author

-number of retweets

Which social factor is more effective to improve relevance?

Factors	RLV-degree
Number of followers	-
Freshness of the tweet	+
The popularity of the author	+
Number of retweet	-
Presence of hashtags	-
Number of mentions	-
Length of the tweet	-
Exact match of terms	+
Presence of URLs in the tweet	+
Language quality	+
Number of replies in the tweet	-
Popularity of the tweet	+

Table 1: Relevance factors [Damak&al.,2013]

Objectives

The goal of this thesis is to:

- Improve the **quality** of results of information retrieval in microblogs.
- Advance the state-of-the-art works by proposing **new solutions** to the short text **representation** and **ranking** problems :

  -Estimate more **accurate representations** of tweets and queries

  -Re-ranking tweets to retrieve **high-quality** content from microblogs

Objectives

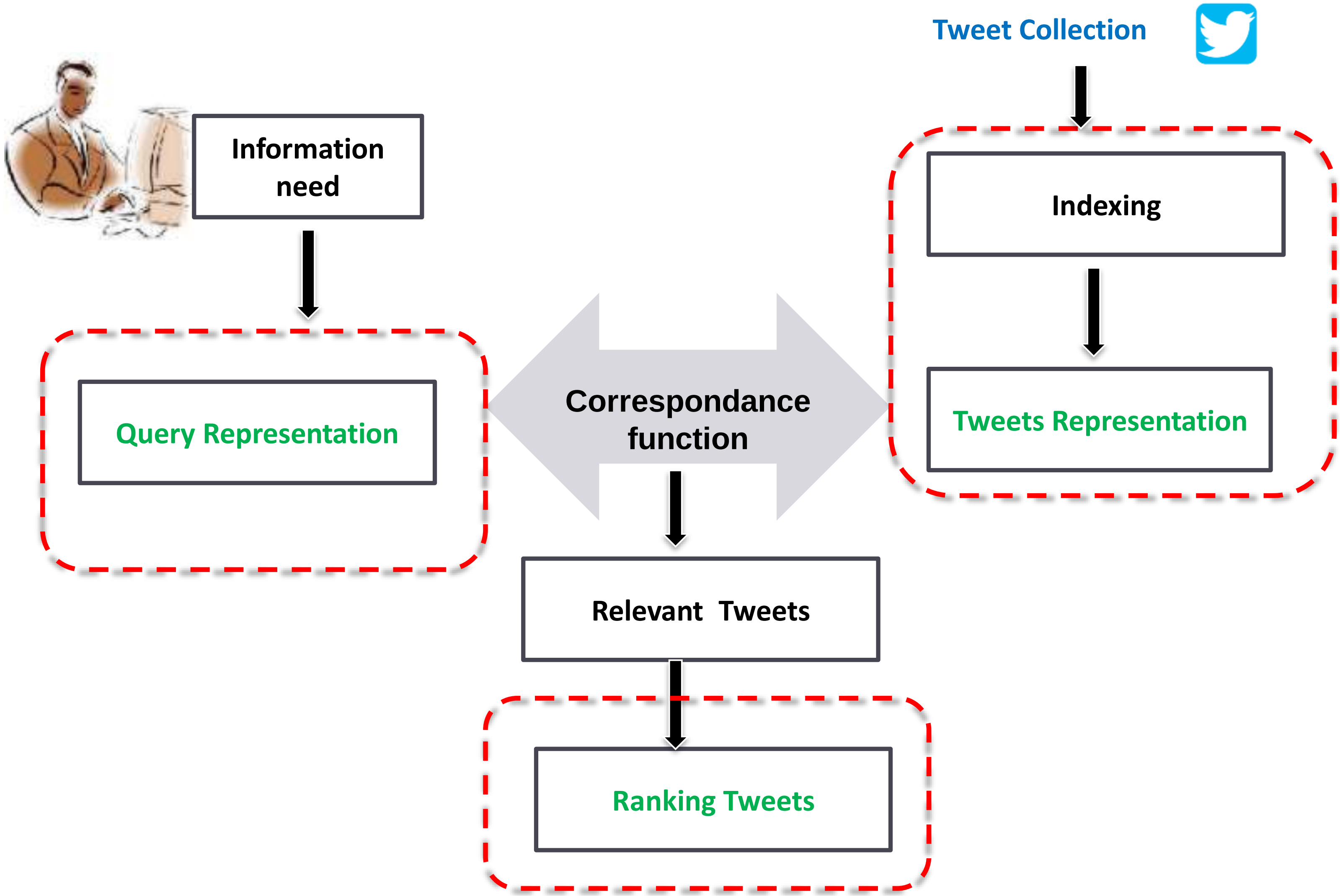


Figure 1: Social information retrieval process

Tweet classification task

Tweet Classification approaches

Learning approaches

•Machine learning algorithms: decision trees , SVM , Naive Bayes

-Large training dataset to train classifier

-If the predefined categories changed: new set of training

-The classification can only be as correct as the labeling of the training set

Semantic orientation approaches

•Knowledge bases

-DBpedia

-Freebase

•Word embedding

-word2vec

Classifier training is required in all classification methods !

Semantic approach for tweet categorization

- A deep enrichment strategy is applied to enrich text tweets with additional semantic concepts from different Knowledge bases (e.g. DBpedia)

- A hard **Word Sense Disambiguation** process that uses a new disambiguation algorithms based on Specification Marks method

- A knowledge-based categorizer called **eXtended WordNet Domain**

➔ A **supervised categorization** which relies only on the ontological knowledge and classifier training is not required.

Model of the Semantic approach for tweet categorization

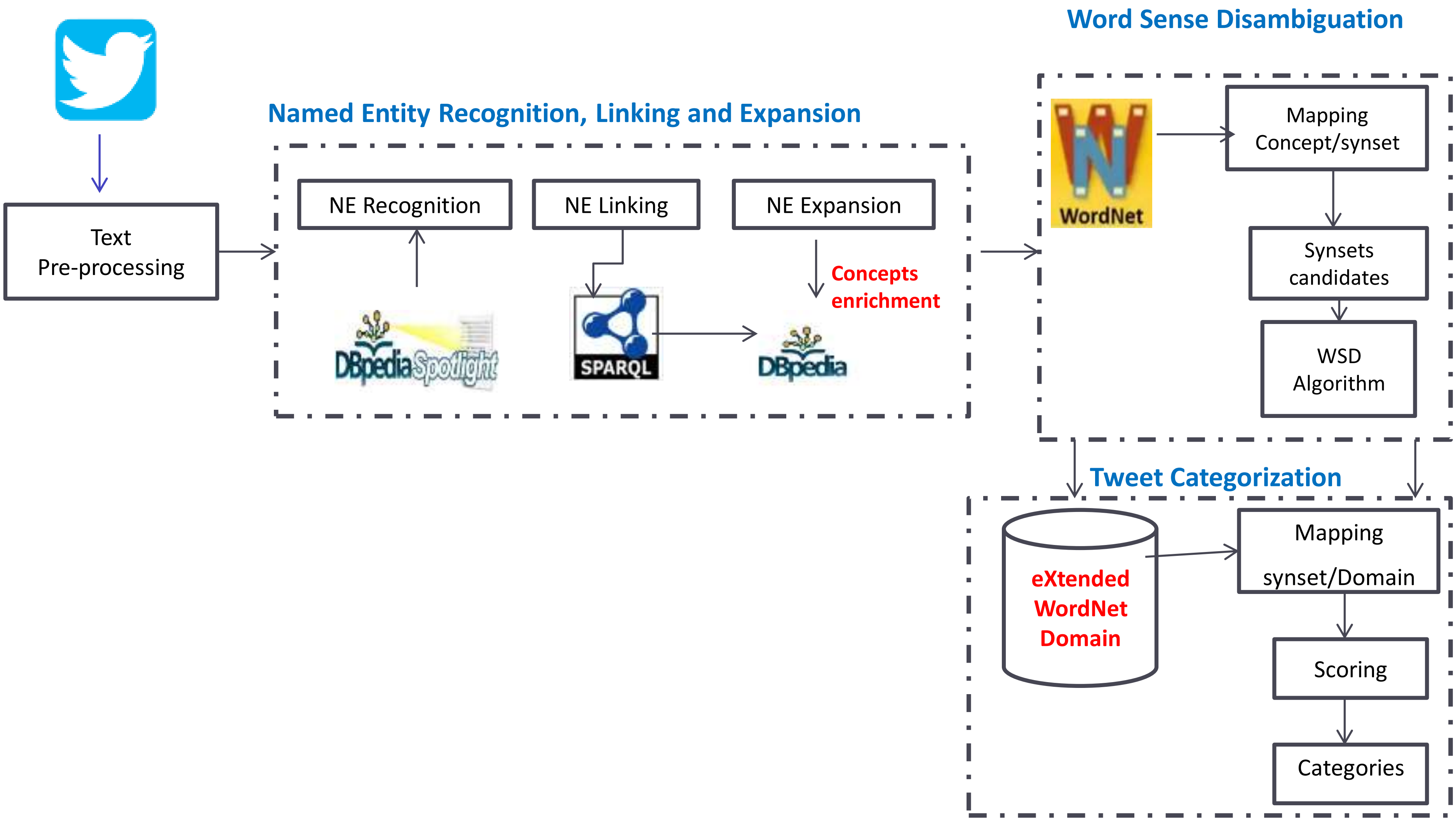


Figure 2: Model of the semantic approach for tweet categorization

Semantic approach for tweet categorization

The semantic approach for tweet categorization can be summarised as follows:

Let

$$D = \{D1, D2, \ldots, Dn\}$$

be the set of **XWND**

Let

$$C = \{c1, c2, \ldots, cm\}$$

be the set of synonymous concepts aggregated in **WordNet** (synsets)

Now, let ***Wi*** *be a word and let*

$$Sense(wi) = \{ci \mid ci \in C\}$$ ***(sense disambiguated)***

with ***Ci*** *being a sense for* ***Wi***

Let

$$T = \{w1, w2, \ldots, ws\}$$

be the whole words in the tweet.

Then, for each sense of word in the tweet **Sense (wi)** , we consider only the domain with the highest **PageRank** weight.

XWND assigns a score to each pre-defined domain annotated score (wi, Dj)

The domain relevance function D* for a word has the following definition:

$$D^* = argmax \sum_{\substack{\forall w_i \in T \\ \forall D_j \in XWND}} score(w_i, D_j)$$

Finally, the tweet is then assigned a label corresponding to the topic (domain)

Evaluation

- Tweet collection which covers 1330 tweets collected via Twitter search API.
- Limited to a six specific topics: Sports, Business, Technology, Entertainment, Politics and Education.
- Only English tweets are included in this evaluation

Features	Accuracy	Recall	Precision	F-measure	Error Rate
Our approach	**91.29%**	**88.25%**	**88.79%**	**88.52%**	**8.71%**
BoE+concepts	87.09%	59.18%	60.96%	60.06%	12.91%
BoS	86.39%	59.79%	59.36%	59.57%	13.61%
BoE+synsets	83.99%	50.17%	51.23%	50.69%	16.01%
BoW	83.61%	50.83%	50.61%	50.72%	16.39%
BoE	81.21%	15.05%	10.54%	8.27%	19.79%

Table 2: Results of tweet categorization

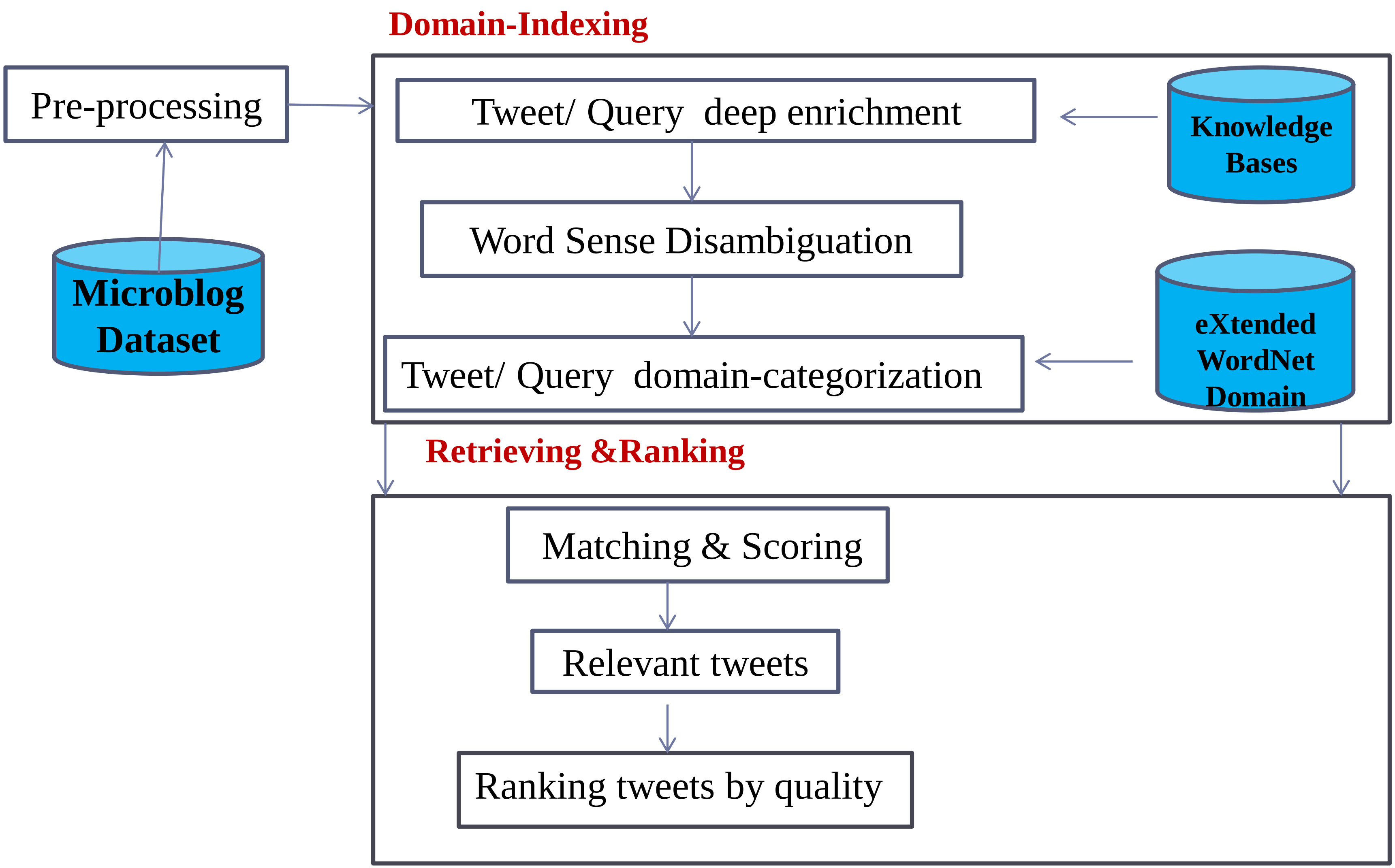


Figure 3: Retrieval model based on domain-specific indexing

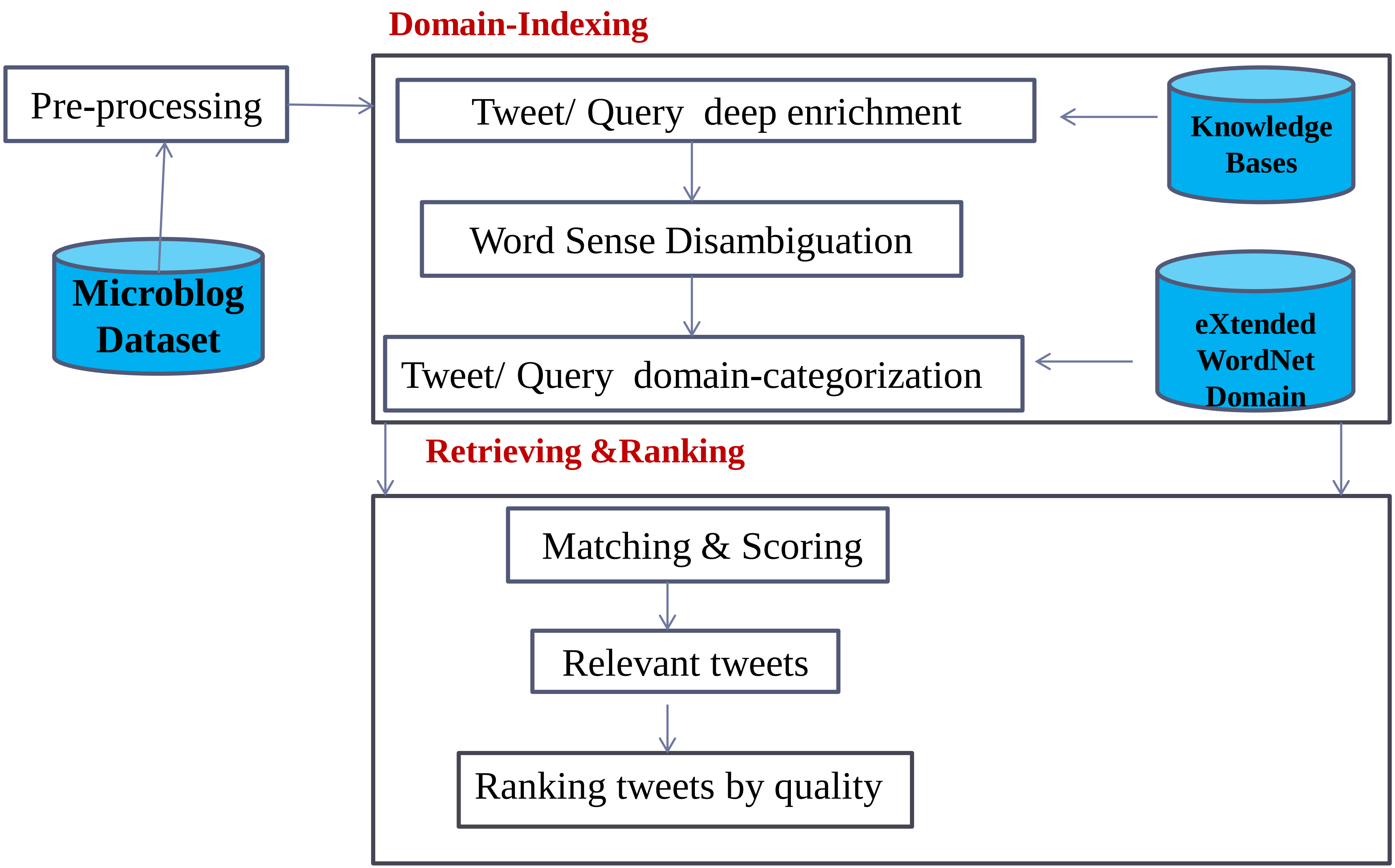


Figure 3: Retrieval model based on domain-specific indexing

Data Collection

- Database from TREC’11 microblog track
- The database consists of 16 million tweets and 49 topics

Evaluation in microblog retrieval

Run	P@5	P@10	P@20	P@30
run-DSI	0.3054	0.3121	0.2972	0.2841
run-KWI -	0.1288	0.1338	0.1321	0.1293

Table 4: Results of retrieval model based on domain-specific indexing

Traditional semantic representation problems

Existing semantic based text representation methods depends on:

-Handcrafted features

-External information sources such as ontologies and knowledge bases

➔**Time-consuming** and **hard hand-engineering**

Need Machine Learning ?

Traditional semantic representation problems

Existing semantic based text representation methods depends on:

-Handcrafted features

-External information sources such as ontologies and knowledge bases

➔**Time-consuming** and **hard hand-engineering**

Need Machine Learning ?

Our improvements of short text representations:

Hybrid Deep Neural Network (HDNN)

- New neural architecture which combines **recurrent neural network** and **feedforward neural network**

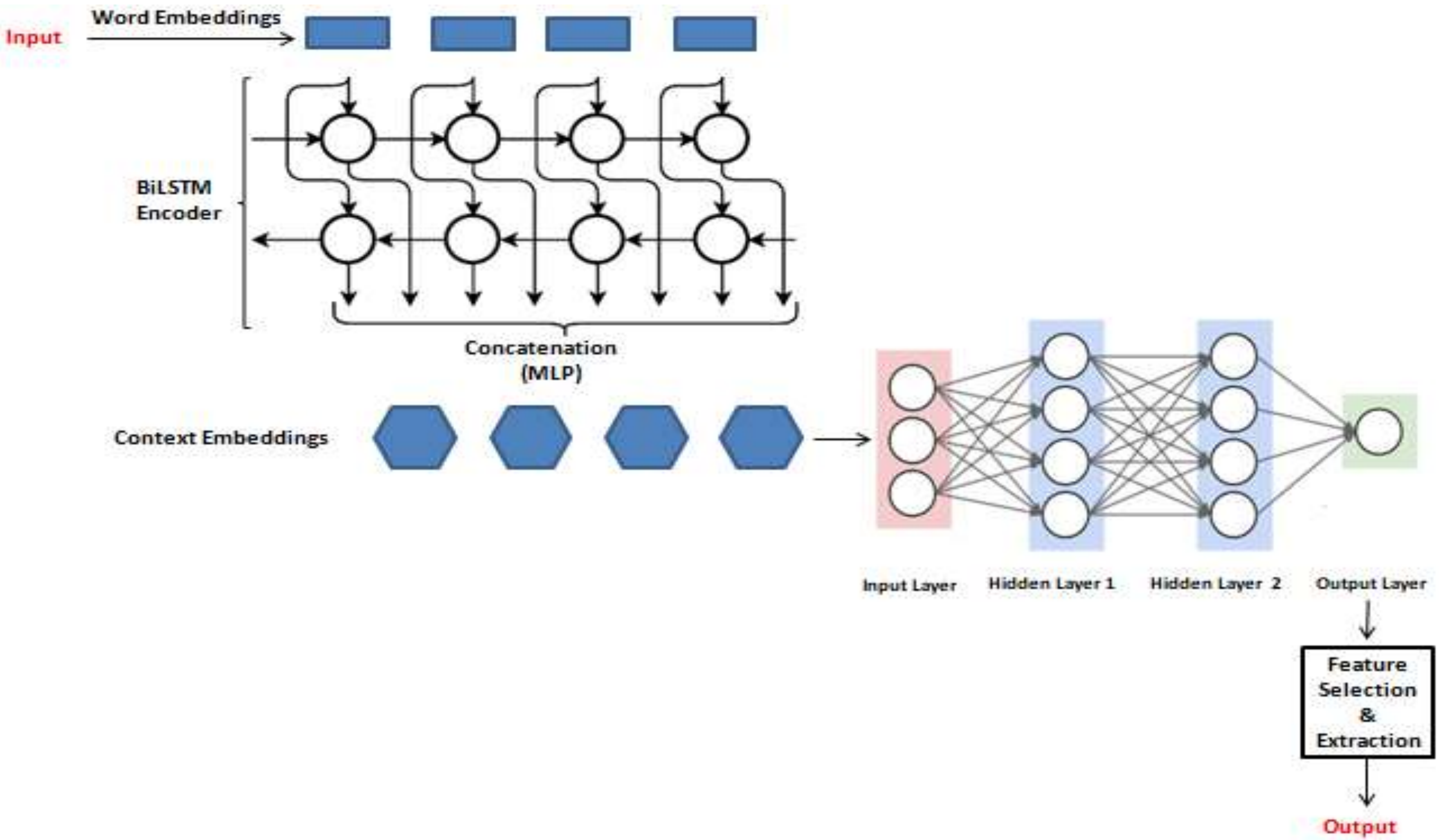
Deep contextualized word representation

- Incorporates character n-grams (FastText) for generating a contextual embedding
- Uses a **bi-directional LSTM**

Hybrid Regularized Autoencoder (HRA)

- Combines **autoencoder** with **Elastic Net regularization** for unsupervised **features selection** and **extraction.**

Hybrid Deep neural network-based representation



Learning features from autoencoder

- Autoencoder is a type of **neural network** that applies back propagation to reconstruct its input data.

- Automatically learns features from unlabeled data by forcing the hidden (encoding) layer to compress the data into a **low-dimensional** representation.

Hybrid regularized Autoencoder

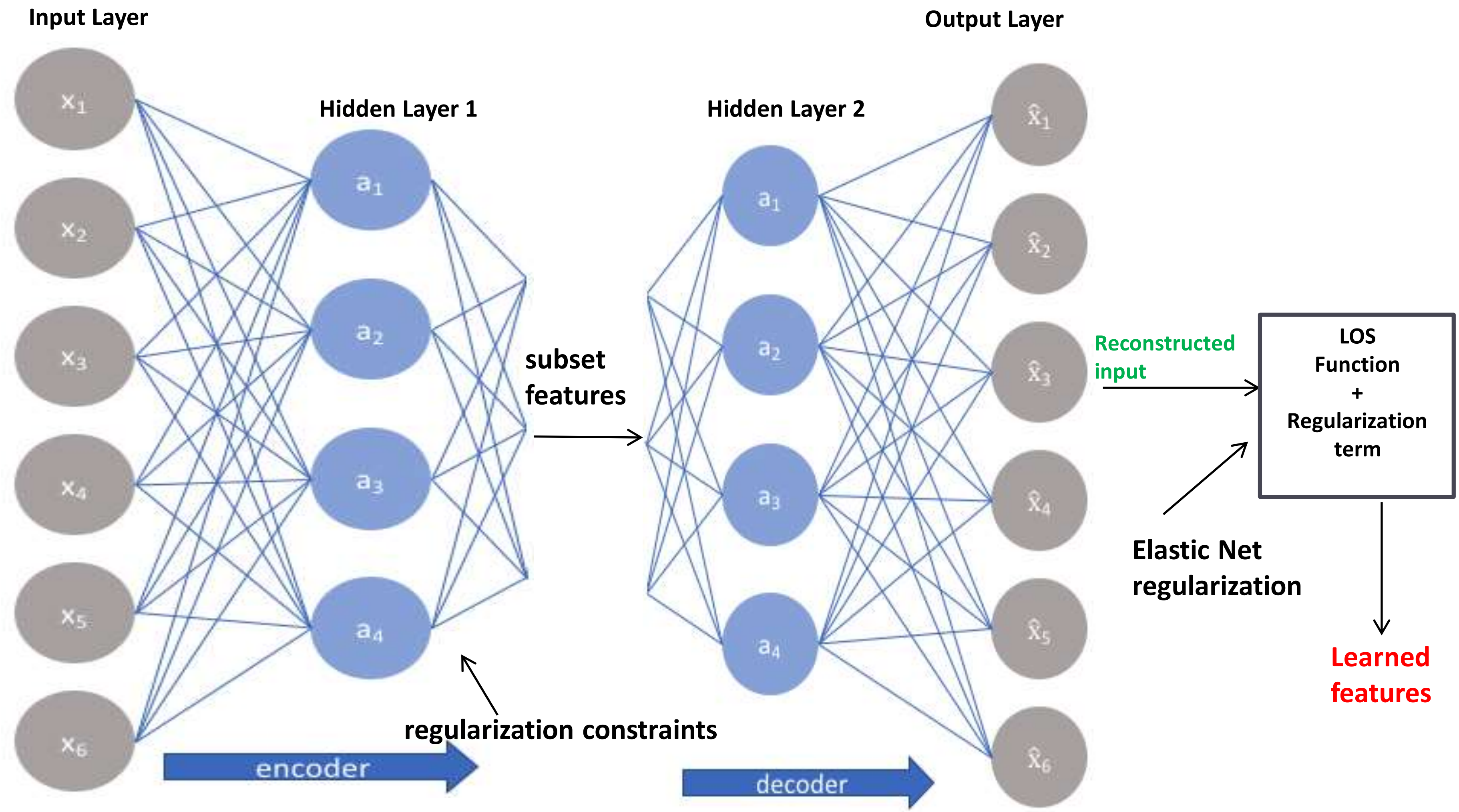


Figure 4 : Hybrid Regularized Autoencoder architecture

Deep Learning representation-based retrieval model

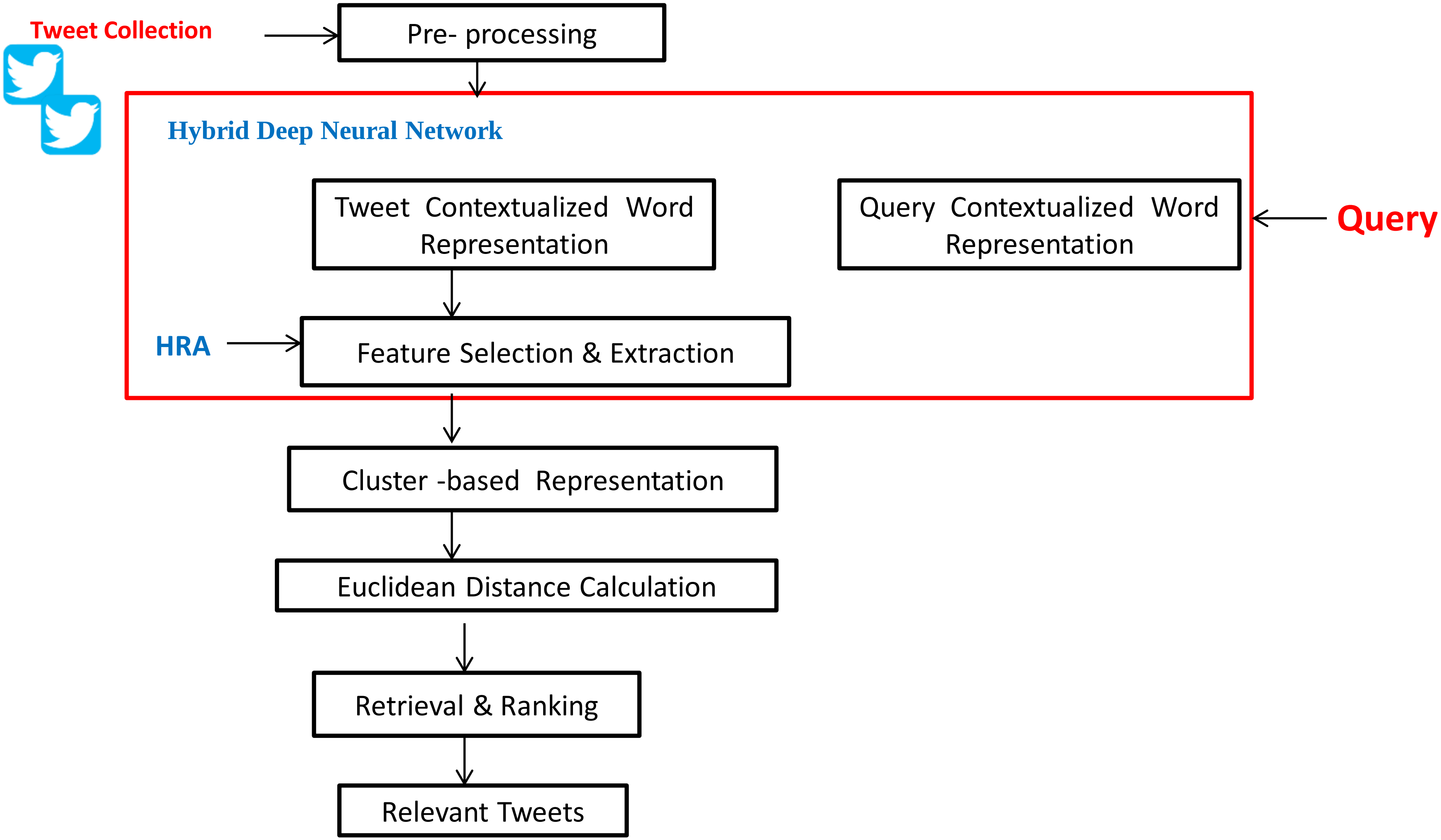


Figure 5: Deep Learning representation-based retrieval model

Run	P@5	P@10	P@20	P@30	MAP
HDNN(ContxW+ HRE)	**0.5817**	**0.5753**	**0.5541**	**0.5468**	**0.5323**
ContxWR+AE	0,5291	0.5110	0.4967	0.4822	0.4708
ContxWR	0,4514	0.4429	0.4284	0.4137	0.4097
FastText	0.4005	0.3843	0.3671	0.3582	0.3454
Word2vec	0.3622	0.3501	0.3363	0.3285	0.3028
GloVe	0,3498	0.3261	0.3033	0.2985	0.2880
LSA	0.2840	0.2724	0.2586	0.2312	0.2059
TF-IDF	0.2086	0.1922	0.1861	0.1677	0.1505
BoW (baseline)	0.1838	0.1738	0.1621	0.1493	0.1234

Table 5: Results of the retrieval process using different representation methods

Model Type	Model	P@30
Our model	HDNN	0.5468
Traditionnal	**BM25 (**TF-IDF)	**0.1293**
	QL-LM (language model)	0.2067
	INDRI-LM (language model)	0.2918
Query expansion	**LCE-QE** (Latent Concept)	**0.4551**
	TM-QE (Text-Mining)	0.2918
	Hybrid-QE	0.3197
	PM-QE (pattern mining)	0.1973
Learned representation	**SA-LM (**Selection Attribute)	**0.3356**
	Auto-LM (autoencoder)	0.1968

Table 6: Comparison with state-of-the-art models

Microbloging ranking problems

Most of the proposed **ranking strategies** :

- provides no guarantee that the most relevant tweets appear on top list
- based on **machine learning algorithms** that depend heavily on **hand-crafted** features (e.g. the number of followers, number of hashtags,etc.),

➔ Feature engineering requiring a lot of **time** and **efforts**

To filter the quality of relevant tweets, we propose:

▪A deep learning ranking approach based on **k-means** clustering to distinguish **high quality** and low quality tweets

•The clustering algorithm is based on learning features from **autoencoder** and **hand-crafted** features from tweets' content and authors' profiles.

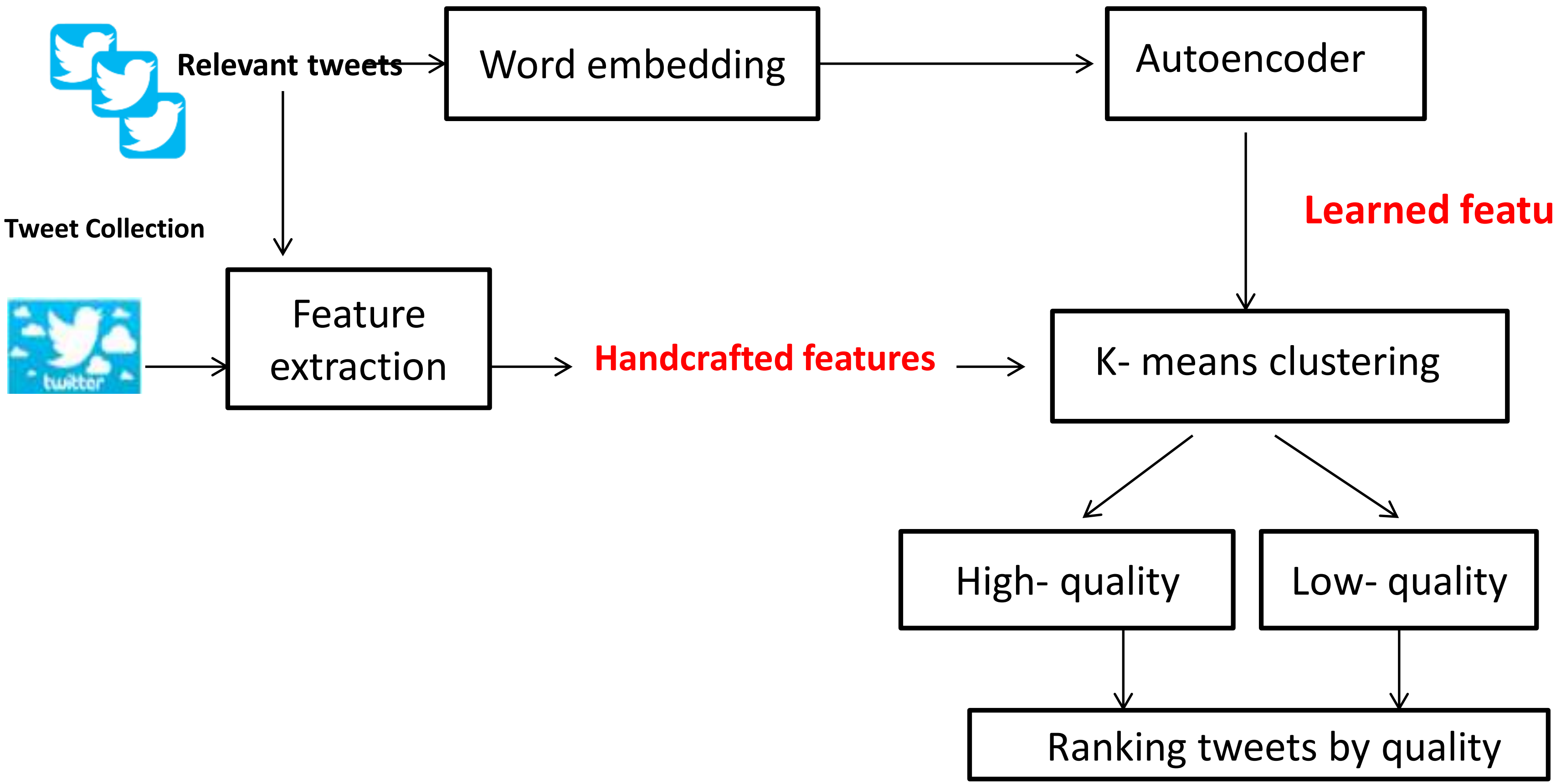


Figure 6: Overall process of re-ranking tweets

Learning features from autoencoder

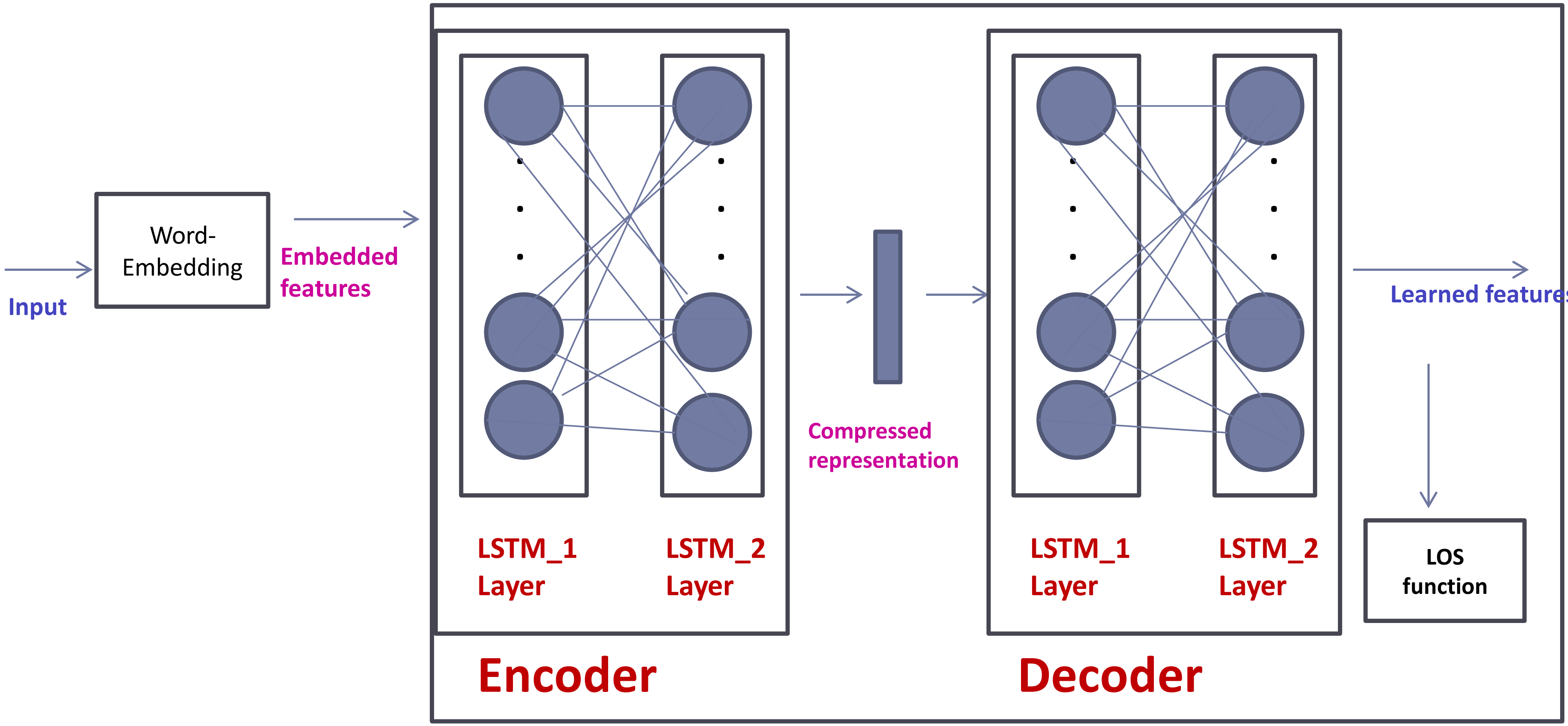


Figure 7: Autoencoder neural network architecture

Defining Handcrafted features

Four types of hand-crafted features were used for distinguishing the tweet's quality content:

- **Structural features** : tweet length, presence of hashtags/named entities
- **Well-formedness features** : spelling / grammar check, number of repeated characters
- **Author profile features**: presence of author profile description
- **Interaction and behavioral features**: number of re-tweets/replies/ mentions

After clustering, we rank tweets in each cluster by measuring the separation distance between the data points and the cluster's centroid using this formula:

$$minD = argmin \sum_{\forall\{X_j \in C_i\}} \|(X_j - cent)^2\|$$

Ranking process

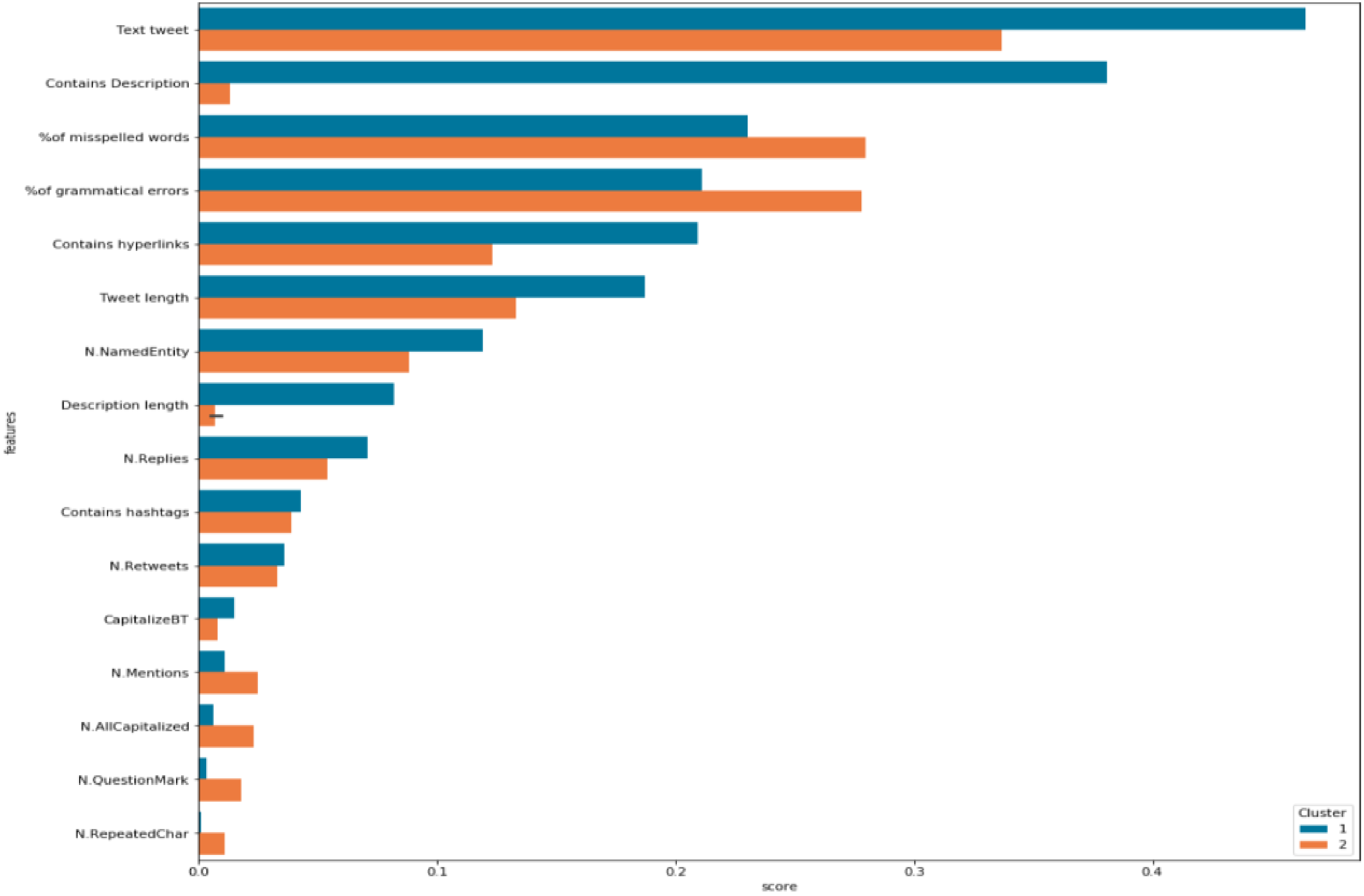


Figure 8: Feature ranking by information gain

Evaluation of the Ranking approach

Run	P@5	P@10	P@20	P@30	MAP	MAP-Gain
Cluster 1 (with LF)	0.2310	0.2267	0.2226	0.1968	0.1881	**81%**
Cluster 2 (with LF)	0.1288	0.1337	0.1371	0.1352	0.1108	7%
run- baseline	0.1288	0.1337	0.1321	0.1293	0.1034	-

Table 7: Ranking results based on k-means clustering with learned features

Evaluation of the Ranking approach

Run	P@5	P@10	P@20	P@30	MAP	MAP-Gain
Cluster 1 (without LF)	0.2145	0.2145	0.2059	0.1882	0.1510	**46%**
Cluster 2 (without LF)	0.1928	0.1928	0.1870	0.1611	0.1356	31%
run- baseline	0.1288	0.1337	0.1321	0.1293	0.1034	-

Table 7: Ranking results based on k-means clustering without learned features

Conclusion

The major contributions of this thesis are :

(1) We proposed a semantic approach for short text representation

It used as a specific domain-indexing technique to improve microblog retrieval

This new indexing technique achieve an improvement of **15%** at **P@30** compared to baseline (Keywords indexing)

Conclusion

(2) We proposed a new representation learning technique which deploys a hybrid neural network architectures:

- The combination of two neural network architectures strongly improves the performance of learning models to extract high-quality features' representations
- This technique achieve an improvement of **42%** at **P@30** compared to state of the art representation techniques

Conclusion

(3) The last contribution consists on a re-ranking approach which aim to retrieve high-quality content from microblogs:

- The integration of the learned features can improve the quality of ranking compared to the use of hand-crafted features only.
- The re-ranking approach achieve a gain of **81%** at **MAP** compared to the reverse chronological order ranking.

Merci pour votre attention

Ibtihel BEN LTAIFA

References

Piyush Bansal, Somay Jain, and Vasudeva Varma. Towards semantic retrieval of hashtags in microblogs. In *Proceedings of the 24th International Conference on World Wide Web, pages 7–8, 2015.*

Paul Ferguson, Neil O'Hare, James Lanagan, Owen Phelan, and Kevin McCarthy. An investigation of term weighting approaches for microblog retrieval. In *European Conference on Information Retrieval, pages 552–* 555. Springer, 2012.

Lin, Y., Li, Y., Xu, W., & Guo, J. (2012, December). Microblog retrieval based on term similarity graph. In *Proceedings of 2012 2nd International Conference on Computer Science and Network Technology* (pp. 1322-1325). IEEE.

Moumita Basu, Anurag Roy, Kripabandhu Ghosh, Somprakash Bandyopadhyay, and Saptarshi Ghosh. A novel word embedding based stemming approach for microblog retrieval during disasters. In *European Conference on Information Retrieval, pages 589–597. Springer, 2017.*

Dwaipayan Roy, Debjyoti Paul, Mandar Mitra, and Utpal Garain. Using word embeddings for automatic query expansion. *arXiv preprint arXiv:1606.07608, 2016.*

References

Fahd Kalloubi et al. Microblog semantic context retrieval system based on linked open data and graph-based theory. *Expert Systems with Applications,* 53:138–148, 2016.

Meriem Amina Zingla, Chiraz Latiri, Philippe Mulhem, Catherine Berrut, and Yahya Slimani. Hybrid query expansion model for text and microblog information retrieval. *Information Retrieval Journal, 21(4):337–367, 2018.*

Archana Godavarthy and Yi Fang. Cross-language microblog retrieval using latent semantic modeling. In *Proceedings of the 2016 ACM International Conference on the Theory of Information Retrieval, pages 303–306, 2016.*

Firas Damak, Karen Pinel-Sauvagnat, Mohand Boughanem, and Guillaume Cabanac. Effectiveness of state-of-the-art features for microblog search. In *Proceedings of the 28th Annual ACM Symposium on Applied Computing,* pages 914–919, 2013.

Arifah Che Alhadi, Thomas Gottron, J´erˆome Kunegis, and Nasir Naveed. Livetweet: Microblog retrieval based on interestingness and an adaptation of the vector space model. In *TREC, 2011.*