International Journal of Science and Research (IJSR)
ISSN: 2319-7064
ResearchGate Impact Factor (2018): 0.28 | SJIF (2019): 7.583

Classification of Arabic News Texts with Fasttext Method

Ozer Celik

[1]Eskisehir Osmangazi University, Department of Mathematics-Computer, Eskisehir, Turkey

Abstract: ***Access to information is getting easier day by day because of the entering our lives of the computer and especially the internet. As the internet access becomes easier and the internet users increase, the amount of data is growing every second. However, in order to access correct information, data must be classified. Classification is the process of separating data according to a certain semantic category. Dividing digital documents into semantic categories significantly affects the availability of the text. In this study, a text classification study was carried out on a data set obtained from different Arabic news sources. Firstly, the news texts were pre-processed and trunked. Pretreated texts were classified by K-Neighbors Classifier, Gaussian Naive Bayes, Multinomial Naive Bayes, Logistic Regression, Random Forest Classifier, Support Vector Classifier and Decision Tree Classifier methods after the FastText method was vectorized. According to the results of the study, the highest success rate was obtained by classification of the text obtained with the FastText vector model with approximately 90.36% with Logistic Regression.***

Keywords: Text Classification, Arabic News, Fasttext

1. Introduction

Access to information is getting easier day by day because of the entering our lives of the computer and especially the internet. However, in order not to get lost among the data produced at a very fast rate and to protect against unwanted data, the obligation to classify the data was born. Text classification has become an important issue, from e-mails to text messages to websites, which we use frequently in our daily lives. In order to reach the desired content and to protect from unwanted data, the texts should be best defined.

The purpose of machine learning is to determine complex problems in information systems and provide them with rational solutions. This shows that machine learning is closely related to fields such as statistics, data mining, pattern recognition, artificial intelligence and theoretical computer science, and requires a multidisciplinary study [1]. Machine learning methods are used effectively for text classification. The purpose of text classification is to automatically split documents into a specific semantic category. Natural language processing (NLP), data mining and machine learning techniques are used together to automatically classify electronic documents. Each document can be uncategorized, in one or more categories. The purpose of machine learning is to learn classifiers from examples that automatically assign categories [2] [3]. In order to classify texts by machine learning, it is necessary to express the texts numerically. Various approaches have been developed for this process. It has been observed that Gensim method, which is one of the methods used to extract vector models of texts, gives better results than FastText library literature review.

1.1. Word Embedding Techniques

Word embedding, also known as word representation, plays a vital role in the creation of continuous word vectors according to the meaning of the word in the document. Word embedding captures both semantic and syntactic information of words and can be used to measure word similarities commonly used in NLP tasks [4].

1.1.1. FastText

Gensim is an open source library for unsupervised subject modeling and natural language processing using modern statistical machine learning. Gensim is implemented in Python and Cython. Gensim is designed to process large text collections using data flow and incremental online algorithms; this makes it different from most other machine learning software packages that only target in-memory processing. Gensim, fastText, word2vec and doc2vec algorithms, as well as flow parallel parallel applications, are hidden semantic analysis (LSA, LSI, SVD), non-negative matrix factorization (NMF), hidden Dirichlet allocation (LDA), Term Frequency Inverse Document Frequency (TF-IDF) and random projections [5]. Some of the new online algorithms in Gensim have also been published in Gensim's creator Radim Řehůřek's Semantic Analysis Scalability in Natural Language Processing in his doctoral thesis. Since 2018, Gensim has been used and quoted in more than 1,400 commercial and academic applications, in various disciplines, from medicine to insurance claims analysis and patent research. The software has been covered in several new articles, podcasts and interviews [6].

As can be understood from the term, TF-IDF calculates the values of each word in a document by the frequency of the word in a particular document with the adverse of the percentage of documents in which the word appears. Basically, TF-IDF works by determining the relative frequency of words in a given document based on the inverse ratio of that word on the entire data set. Intuitively, this calculation determines how relevant a particular word is to a particular document. Words common to a single or small group of documents tend to have higher TF-IDF numbers than general words [7].

$$w_{i,j} = tf_{i,j} \times log\frac{N}{df_i} \quad (1)$$

$tf_{i,j}$ = frequency of the word i in the j document
df_i = number of documents containing the word i
N = total number of documents

Word2Vec was proposed by Mikolov et al. In 2013 [8]. This method establishes a close relationship between the word and its neighbors in a given window size and positions the words that are close to meaning in a way that they are close to each other in the vector space. It uses two different learning architectures to establish meaning relations.

The first is the Continous Bag of Words (CBoW) architecture. In this method, the word in the window center is tried to be guessed by looking at the neighbors of the word as close as the window size.

The other method is Skip-Gram architecture. This method works similarly to CBoW. However, unlike CBoW, Skip-Gram predicts the neighbors of the word located in the center of the window. The advantage of this method is that it can capture multiple meanings of words that can have different meanings.

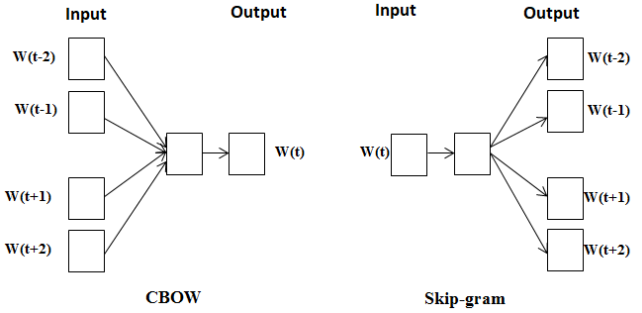

Figure 1: CBoW and Skip-Gram model architecture

FastText is a Word2Vec based model developed by Facebook in 2016. The difference of this method from Word2Vec is that words are divided into n-grams. Thus, the meaning closeness that cannot be captured with Word2Vec can be captured with this method [9].

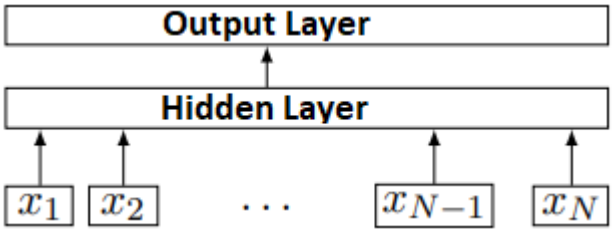

Figure 2: FastText model architecture

1.2. Statistical analysis

After applying machine learning techniques to the data set used in the research, the accuracy rates were calculated using the confusion matrix. Confusion matrix is the table that gives the number of correct and incorrectly classified data groups in a dataset [10].

In our study, General Accuracy Rate (OACC), which is a success assessment method widely used in multi-class data, was used [11].

Table 1: Multiclass Confusion matrix

Data Set		*Actual Durum*			*Total*
		Class 1	*Class 2*	*Class 3*	
Predict	*Class 1*	a	b	c	j
	Class 2	d	e	f	k
	Class 3	g	h	i	l
Total		m	n	o	N

The success scores of the multi-category models are calculated with the help of the confusion matrix (Table 2). Success measures and formulas used in our study and calculated with the help of confusion matrix;

Table 2: Confusion matrix success measures and formulas

N = (a + b + c + d + e + f + g + h + i)
j = a + b + c
k = d + e + f
l = g + h + i
m = a + d + g
n = b + e + h
o = c + f + i
OACC = (a + e + i) / N
Precision1 = a / j
Precision2 = e / k
Precision3 = i / l
Recall1 = a / m
Recall2 = e / n
Recall3 = i / o
Specificity1 = (e + f + h + i) / (k + l)
Specificity2 = (a + c + g + i) / (j + l)
Specificity3 = (a + b + d + e) / (j + k)

In all analyzes and operations, a computer with quad core Intel Skylake Core i5-6500 CPU 3.2 GHz 6 MB cache and 8 GB 2400 MHz DDR4 Ram memory was used.

1.3. Literature Review

Abdul-Mageed et al. suggested a machine learning approach to perform a subjectivity and sentiment classification at the sentence level. They first collected 11,918 sentences of different types from a news site and created a dataset by manually tagging them. Classification is carried out in two stages. At the first stage, they made a distinction between a subjective and objective text (ie Subjectivity Classification). In the second stage, they explored a difference between a positive and negative emotion (ie, Emotion Classification). In this study, Abdul-Mageed et al. It used SVM as a learning algorithm with language-specific and general features. Language-independent features include n-gram, domain name, unique, and polarity dictionary features. Arabic-specific features have been added to investigate the impact of morphological information on performance. The findings have shown that using POS labeling and lemmas or lexemes to extract basic forms of words has a positive effect on subjectivity and emotional classification [12, 13].

Nabil et al. they presented a 4-way (objective, subjective negative, subjective positive and subjective mixed) classification of emotions that classifies texts into four

classes: The datasets contain 10,006 Arabic Tweets that are manually tagged using the Amazon Mechanical Turk (AMT) service. They applied various machine learning algorithms to datasets. However, the use of n-grams as uniform properties in versatile classification did not yield good results [14].

El-Makky et al. Combined the Emotional Orientation algorithms with a machine learning classifier to suggest a hybrid approach. For each document in the Twitter dataset, they used a dictionary-based approach to calculate Emotional Orientation scores. These scores are integrated with different features such as unigrams, language-independent features, Tweet-specific features, and body polarity features to create an input feature vector for the SVM classifier. This combination of the Machine Learning classification approach and the dictionary-based approach led to slightly better results than an approach result (accuracy 84%) [15].

El-Baltagy et al. Egypt proposed a dictionary-based approach to create a moral classification for Arabic texts. After creating a dictionary with 4,392 terms, the authors used two datasets (500 tweets' Twitter dataset and a Dostour dataset from 100 comments on the web) to evaluate an uncensored classification algorithm. First, it adds weight to negative and positive terms and calculates a score for each document. The second algorithm assigns a positive and negative weight to each term in the dictionary and calculates positive and negative scores for each document. The authors achieved good results using two algorithms (83.8% accuracy) in the Twitter dataset [16].

2. Materials and Methods

The data set was compiled from the National Iraqi News Agency news site using Python's BeautifulSoup library. There are 2599 news articles in economy, security, sports, politics and press categories. The categories were determined based on the categories determined by news sites.

2.1. Data Pre-Processing

There are many characters / words in the raw data set that are not important for machine learning. In order to get more precise results, it is necessary to clear the texts from these characters / words. For this, the following operations were applied on the raw data.

- Unnecessary characters and numerical expressions such as operators, punctuation marks are cleared.
- By converting all the words to lowercase, the same words can be perceived by the machine as different words.
- Turkish ineffective words (stopwords) have been deleted.
- By finding the roots of the words, it is aimed to perceive only the words with different suffixes as the same word.

2.2. Word Embedding Process

FastText method was used to obtain the vector model of the data set. Words that have passed at least 5 times in the whole data set have been evaluated. In FastText method, minimum n-gram 2 is determined as maximum n-gram 10. Default values are used for values other than this.

3. Results and Discussion

In this study, Fasttext method was applied to obtain the vector model of the words in the data set. The data set, which has been pre-processed and removed from the vector model, is divided into 70-30% training and test documents. The word vectors obtained were trained and compared with Logistic Regression, Multinomial Naive Bayes, Support Vector Classifier, K-Neighbors Classifier, Gaussian Naive Bayes, XGB Classifier, MLP Classifier, Decision Tree Classifier and Random Forest Classifier, and then compared.

The success rates obtained from the training of the data, whose vector model was created with FastText, are given in Table 3.

Table 3: Accuracy rates

	FastText	
	Accuracy	*F-Score*
Logistic Regression	%90.359	%90.2818
Multinomial Naive Bayes	%77.538	%78.465
Support Vector Classifier	%90.154	%90.100
K-Neighbors Classifier	%77.949	%77.646
Gaussian Naive Bayes	%80.513	%80.782
XGB Classifier	%88.513	%88.513
MLP Classifier	%89.846	%89.805
Decision Tree Classifier	%68.497	%68.497
Random Forest Classifier	%81.641	%81.710

The accuracy rates achieved by category are given in Table 4.

Table 4: The accuracy rates by categories

%	*FastText*								
	Logistic Regression	*Multinomial Naive Bayes*	*Support Vector Classifier*	*K-Neighbors Classifier*	*Gaussian Naive Bayes*	*XGB Classifier*	*MLP Classifier*	*Decision Tree Classifier*	*Random Forest Classifier*
Economi	94.36	76.41	94.87	68.21	77.95	90.77	93.33	69.74	84.10
Security	94.36	85.13	92.31	86.15	86.15	92.82	90.77	72.82	83.08
Sport	97.44	84.10	98.46	94.36	86.67	96.41	96.92	86.15	94.87

Politics	74.95	76.41	77.44	45.12	66.67	79.49	78.46	52.82	71.28
Press	87.69	65.64	87.69	95.90	85.13	83.08	89.74	61.54	74.87

Table 5: Confusion Matrix of the Logistic Regression method that gives the best accuracy rates

	Economi	*Security*	*Sport*	*Politics*	*Press*
Economi	184	1	0	8	2
Security	2	184	2	7	0
Sport	1	0	190	1	3
Politics	11	25	3	152	4
Press	6	5	2	11	171

4. Conclusion

As a result of the study, it has been observed that the classification performance of Logistic Regression and Support Vector Classifier algorithms of the vector models obtained with FastText are very close to each other, but the most successful rate is 90.36% with the Logistic Regression model. The most accurately predicted category was the sports category across vector model methods and estimation algorithms.

Acknowledgement

Thanks so much to Abdullah Algumar whose native language is Arabic, for evaluation of Arabic Linguistic.

References

[1] Vapnik V. The nature of statistical learning theory. Springer, 2nd edition, 2013; New York, USA. pp: 32-40.
[2] Joachims, T. (1999, June). Transductive inference for text classification using support vector machines. In Icml (Vol. 99, pp. 200-209).
[3] Khan, Aurangzeb, et al. "A review of machine learning algorithms for text-documents classification." Journal of advances in information technology 1.1 (2010): 4-20.
[4] Liu, Y., Liu, Z., Chua, T. S., & Sun, M. (2015, February). Topical word embeddings. In Twenty-Ninth AAAI Conference on Artificial Intelligence.
[5] Okumura, H., Miki, K., Misawa, S., Sakamoto, K., Sakamoto, T., & Yoshida, S. (1989). Observation of direct band gap properties in gensim strained-layer superlattices. Japanese journal of applied physics, 28(11A), L1893.
[6] Rehurek, R. (2011). Scalability of semantic analysis in natural language processing. Masarykova univerzita, Fakulta informatiky.
[7] Ramos, J. Using tf-idf to determine word relevance in document queries. In Proceedings of the first instructional conference on machine learning 2003; (Vol. 242, pp. 133-142).
[8] Mikolov T, Chen K, Corrado G, Dean J. (2013), "Efficient estimation of word representations in vector space". Proceedings of Workshop at ICLR. Scottsdale, Arizona 2-4 Mayıs 2013.
[9] Joulin, A., Grave, E., Bojanowski, P., & Mikolov, T. (2016). Bag of tricks for efficient text classification. arXiv preprint arXiv:1607.01759
[10] Osmanoğlu, U. Ö., Atak, O. N., Çağlar, K., Kayhan, H., & Can, T. C. Sentiment Analysis for Distance Education Course Materials: A Machine Learning Approach. Journal of Educational Technology and Online Learning, 3(1), 31-48.
[11] https://www.researchgate.net/publication/310799885_Generalized_Confusion_Matrix_for_Mult
[12] iple_Classes
[13] M. Abdul-Mageed, M.T. Diab, M. Korayem, Subjectivity and sentiment analysis of modern standard Arabic, Presented at the Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2, 2011, pp. 587–591.
[14] Abdul-Mageed M, Diab M, Kübler S. SAMAR: subjectivity and sentiment analysis for Arabic social media. Comput. Speech Lang. 2014;28(1):20–37.
[15] M. Nabil, M. Aly, A.F. Atiya, ASTD: Arabic sentiment tweets dataset, Presented at the Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pp. 2515–2519.
[16] N. El-Makky et al., Sentiment analysis of colloquial Arabic tweets, 2015.
[17] S.R. El-Beltagy, A. Ali, Open issues in the sentiment analysis of Arabic social media: a case study, Presented at the 9th International Conference on Innovations in information technology (iit), 2013, pp. 215–220.

Author Profile

Ozer Celik received his B.C. degree in Computer Engineering, Electric & Electronic Engineering Department (Double Major), M.S degree in Electric & Electronic Engineering Department and PhD degree in Mathematics-Computer Department at Eskişehir Osmangazi University. Dr. Celik is lecturer in Department of Mathematics-Computer Science. His interest areas are artificial intelligence, machine learning, deep learning and mobile applications.