International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

An Active Learning Based Emoji Prediction Method in Turkish

Emrah Inan[1,*]

Submitted: 19/08/2019 Accepted: 25/02/2020

Abstract: Emoji usage has become a standard in social media platforms since it can condense feelings beyond short textual information. Recent advances in machine learning enable to write short messages with automatically detected emojis. However, the prediction of emojis for the given short message can be complicated, inasmuch as users can interpret different meanings beyond the intent of their designers. Therefore, an automatic extraction strategy of training samples cannot be convenient from the large volumes of unlabelled tweets. In this paper, we present an active learning method to evaluate the emoji prediction of a tweet with a limited number of labelled Turkish emoji dataset. To simulate a human-machine collaborative learning method, we train an initial classifier with this dataset and then we update the classifier by filtering related samples out from the large pool of unlabelled data. In the evaluation, we split 25% randomly selected tweets combined with only one emoji from the generated dataset as a test case. Our active learning method has achieved 0.901 F1 score and outperforms other baseline supervised learning methods.

Keywords: *Active Learning, Emoji Prediction, Turkish Emoji Dataset*

1. Introduction

Emoji is a graphical expression of emotions which enhances the meaning of a short text message [1]. Emoji were first used by several Japanese mobile operators as a message feature. Then, Unicode Consortium (https://unicode.org/emoji/charts/full-emoji-list.html) presents an emoji list in social media as a dictionary. In 2016, Unicode version 9.0 includes 1126 emoji list and nowadays it reaches to a total number of 1719 emojis. In this list, each emoji has both code and name in Unicode Common Locale Data Repository (CLDR). Therefore, it provides a standard for better taxonomy classification of multiple class and label tasks in sentiment analysis and emoji prediction.

The main difference between multi-class and multi-label classification tasks is the number of assigned labels for each sample. Multi-class classification task assumes that each sample can be assigned to one and only one label and each label are mutually exclusive. On the other hand, multi-label classification task can be considered as a more improved version of the multi-class task. It assigns to each sample a set of target labels which are not mutually exclusive, such as movies can be classified into one or more genres. In our study, we perform a multi-class classification task in which a tweet can be assigned only one emoji.

To label texts with emojis there exist lexicons for multi-class and multi-label classification tasks. Novak et al. [2] produce a lexicon including negative, positive and neutral classes of 1.6 million tweets collected in 13 different European languages. Einer et al. [3] represent all Unicode emojis which are learned from their descriptions as a pre-trained vector space model called as emoji2vec in the Unicode emoji standard. However, it is a very time-consuming operation to label the texts with the corresponding emojis. The prediction of emojis for the given text can be complicated, inasmuch as users can interpret different meanings beyond the intent of their designers.

To overcome this problem, Active Learning aims to perform better with less labelled data and it can improve the performance by selecting the label assignment of data in the unlabelled dataset [4]. Hence, the labelling cost can be reduced through an iterative process.

Each iteration of Active Learning methods consists of a query strategy and it chooses the most informative instances from the unlabelled dataset. To label the selected instances, an oracle is employed to label these instances and then add them to the labelled dataset for inducing the prediction task. In this study, we perform an emoji prediction task based on Active Learning. To evaluate the prediction task, we leverage a text with a limited number of tweets labelled with emoji in Turkish. To simulate a human-machine collaborative learning method, we train an initial classifier with this dataset and then we update the classifier by filtering related samples out from the large pool of unlabelled data.

The remainder of this paper is laid out as follows. Section 2 gives an overview of related work. In Section 3, the proposed method for the emoji prediction using Active Learning method is explained in detail. The experiments are shown for the proposed method and other techniques on the generated evaluation dataset in Section 4. We conclude our study and highlight the research questions in Section 5.

2. Related Work

Recent studies such as sentiment analysis, emotion classification and irony detection have concentrated on the usage of emojis to alleviate the problems. In English, there are various studies considering only the emoji prediction task rather than using emojis as a subtask to improve the performance of sentiment analysis or polarity detection.

Barbieri et al. [5] present an emoji prediction model from tweets and train this model based on Long Short-Term Memory networks. The evaluation results denote that their model outperforms Bag-of-Words (BoW) and Skip-Gram vector average baselines. Li et al.

[1]*National Centre for Text Mining, School of Computer Science, The University of Manchester, United Kingdom*
ORCID ID: 0000-0002-1229-6895
** Corresponding Author Email: emrahinan@gmail.com*

[6] provide a model based on Convolutional Neural Networks to learn emoji embeddings and then the model uses these embeddings for the emoji prediction task.

In Turkish, emojis and emoticons are mostly used as a subtask for the improvement of a sentiment analysis method. For instance, Coban et al. [7] leverage emoticons to address the sentiment analysis problem and uses positive and negative emoticons to assign labels for the sentiments. They use Support Vector Machine (SVM), Naive Bayes, Multinomial Naive Bayes (MNB) and k-nearest neighbors algorithms that are based on BoW and n-gram vector representations. Shiha and Ayvaz [8] investigate the impact of emojis in the sentiment analysis problem and they analyze some global events including "The New Year's Eve" and "Istanbul Attack" as positive and negative events, respectively. Yurtoz and Parlak [9] perform sentiment analysis using SVM and MNB on the train and test sets gathered from Turkish tweets. They grouped emojis into positive and negative emotions.

Velioglu et al. [10] perform a sentiment analysis task and also give a direction to the emoji prediction task by using Naive Bayes, Logistic Regression, SVM and Decision Trees on BoW and fastText vector representations. They also grouped these emoji labelled tweets in positive, negative and neutral emotions types. We differ from this study by using Active Learning method for the only emoji prediction task. To achieve this task, we also generate an emoji labelled dataset for train and test phases in Turkish.

3. Method

Active Learning interests in learning accurate classifiers by selecting which samples are labelled and then it tends to reduce the labelling costs to train an accurate method [4]. There are many data mining and machine learning tools such as WEKA and Scikit-learn for examining current algorithms and evaluating their performance. However, these tools are concentrated on mainly Supervised and Unsupervised Learning tasks. Recently, there exist Active Learning libraries such as LibAct [11] and JCLAL [12] that implement several Active Learning strategies presented in the literature.

In this study, our emoji prediction method requires a multi-class learning algorithm as a subset of multi-label learning. For this reason, we employ JCLAL framework which offers several Active Learning strategies for multi-label learning tasks by using MULAN [13]. This framework is also written in Java and open source library for the development of Active Learning methods.

Figure 1 illustrates an overview of the proposed method. In the first step, we train our method with a limited number of instances from the labelled emoji dataset. In the Active Learning process, we employ the Sequential Minimal Optimization (SMO) [14] method as a base classifier and it is an improved version of Support Vector Machines. We use an adaptation [15] of SMO which deals with the multi-class classifier as pairwise classification tasks of this binary method in Weka [16].

The second step focuses on the query strategy in the Active Learning process and we set the query strategy as Binary Minimum (BinMin) [17] from the multi-label strategies. We transform the multi-label strategy into a multi-class classification by assigning only one label for each tweet. Brinker [17] proposed a generalization strategy for the pool-based active learning and BinMin strategy is based on the common one-versus-all binary decomposition scheme.

In the third and fourth steps, BinMin strategy selects the most critical samples and then show these samples to the simulated oracle, respectively. The last step deals with increasing the size of the labelled dataset with the selected and labelled samples.

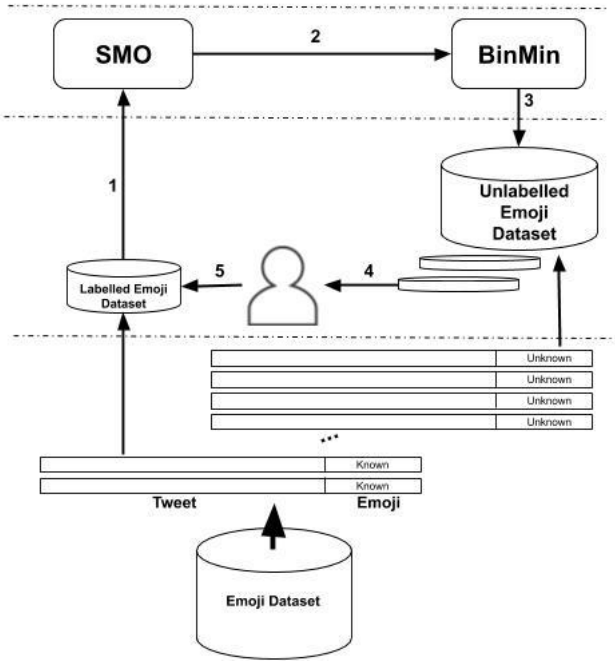

Fig 1. General structure of the proposed method

3.1. SMO Base Classifier

To train the SVM model, a very large quadratic programming (QP) problem arises. To solve this problem, the SMO algorithm split it into a series of smallest possible QP problems. The SMO presents an iterative algorithm for solving these small problems analytically. As a mathematical optimization, Lagrange multipliers are used for detecting the local maxima and minima of a function subject to equality constraint. Then, linear equality constraints are solved as

$$0 \leq \varphi_1, \varphi_2 \leq S, \quad (1)$$

$$y_1\varphi_1 + y_2\varphi_2 = k \quad (2)$$

where S is an SVM hyperparameter, k is the negative of the sum for the multiplication of binary labels y_1 and y_2 and multipliers φ_1 and φ_2 in the equality constraint. In our study, we assign emojis as labels to tweets and extract sample labelled dataset including tweet and known emoji label from the large Emoji dataset as denoted in Figure 1. Then, we train the SMO model to predict labels for the selected tweets supplied by the BinMin query strategy. The simulated oracle adds the most convenient samples which contain tweets and their predicted labels to the current labelled dataset.

3.2. BinMin Query Strategy

In this study, we select the BinMin method as the query strategy which is the most related to the stated problem. The BinMin method selects the optimal unlabeled sample as

$$argmin_{x \in U} = (min_{i=1,2,\ldots,d} \mid (f^i(\mathrm{x})\mid) \quad (3)$$

where ***argmin*** is the notation of worst-case for the unlabeled dataset ***U*** that includes sample ***x***. The minimum absolute distance $min_{i=1,2,\ldots,d}$ is then evaluated among the binary classifier $f^i(\mathrm{x})$ on the binary problem associated with class ***i*** from a set of ***d*** version spaces. Hence, we leverage SMO as the binary classifier and it selects unlabeled samples (Step 3 in Fig. 1) concerning the most uncertain label (Step 4 in Fig. 1). We iterate these steps until reaching the highest performance scores. In the following section,

we give details about the experimental setup and performance of the proposed method.

4. Evaluation

In the evaluation, we first generate a dataset including Turkish tweets with their emojis as both training and test cases. Then, we compare our active learning method with three baselines.

4.1. Dataset

To compare the proposed method with the classical supervised learning methods, we employ the Twitter4J library (http://twitter4j.org/en/} to retrieve 4358 tweets. Tweets were posted between May 2019 and August 2019 geo-localized in Turkey. To preprocess of the generated dataset, all hyperlinks, mentions, and hashtags from each tweet are removed and lowercased all textual content to reduce noise [18]. Also, we only focus on emojis and remove emoticons such as :) and :(in the preprocessing step and we collect tweets including more than 50 characters after this step.

To produce the dataset, we selected tweets which consist of one and only one emoji of the 10 most frequent emojis that is denoted in Table 1.

Table 1. Emojis in the generated dataset

Emoji code	Short Name	Frequency
U+1F600	grinning face	482
U+1F602	face with tears of joy	464
U+1F60D	smiling face with heart-eyes	452
U+1F60E	smiling face with sunglasses	444
U+2764	red heart	432
U+1F4AA	flexed biceps	425
U+1F64F	folded hands	421
U+1F615	confused face	418
U+1F48B	kiss mark	414
U+1F620	angry face	406

Table 1 represents emoji codes and short names for the selected emojis from the Unicode Emoji Charts. In this study, we select the 10 most frequent emojis from the list to the emoji prediction task. Also, we demonstrate the number of tweets where each emoji appears to experiment with the effect of the frequency for the prediction performance.

In the experimental setup, we employ the 10 most frequent tweets with their emoji labels as for the evaluate the proposed method for this 10-class prediction performance. As denoted in the table, the frequency of these emojis is almost well balanced between 406 and 482. For this balanced data set, our text cleaning step contains HTML decoding, stop words, punctuation and non-characters elimination.

Turkish tweets tend to be informal and free writing style and are not generally in the canonical sentence structure [19]. For this reason, we apply Turkish morphological analyzer [20] to perform Turkish deasciifier and spell checker. After that, we split them into tokens and remove stop words. Finally, we lemmatize each token by using Python NLTK Snowball Stemmer in Turkish.

4.2. Evaluation Measures

To compare the performance of the proposed method with three baselines supervised learning methods, we use Precision, Recall and F-1 measure.

$$\text{Precision} = \frac{TP}{TP+FP} \quad (4)$$

Precision is the ratio of the true positive (TP) classifications against the summation of the TP and false positives (FP) as illustrated in Equation 1.

$$\text{Recall} = \frac{TP}{TP+FN} \quad (5)$$

Recall is the proportion of the true positives (TP) against the TP and false negatives (FN) as denoted in Equation 2.

$$\text{F1} = \frac{2*Precision*Recall}{Precision+Recall} \quad (6)$$

F1-score has a range between 0 and 1. Equation 3 shows that it is the harmonic mean of precision and recall values.

$$\text{Acc.} = \frac{TP+TN}{TP+TN+FP+FN} \quad (7)$$

Finally, Accuracy (Acc.) is the measure of all the correctly classified instances.

4.3. Results

To prepare the generated dataset for the evaluation task, we remove each emoji from the tokens in the given tweet and assign it as a label both for training and testing sets. The main task for the proposed method is to predict the single emoji that appears in the given tweet.

We select Linear SVM, Multinomial Logistic Regression (MLR) and MNB as three baseline methods by using in Scikit-Learn tool (http://scikit-learn.org/stable/index.html).

SVM is a method for the classification of both linear and non-linear data. It leverages a non-linear mapping to convert the original training data to a higher dimension. Thanks to this new dimension, it searches for the linear optimal separating hyperplane or decision boundary that separates the samples of one class from other classes. The SVM finds the hyperplane using support vectors and margins [21].

MLR is a supervised learning method that generalizes logistic regression to multi-class problems. MLR is a predictive analysis and is used to explain the relationship between one nominal dependent variable and one or more independent variables [22].

MBN is a probabilistic supervised learning method and is a specific instance of a Naive Bayes classifier which uses a multinomial distribution for each of the features. Due to the dependency of on Bayes' theorem, it has an assumption of conditional independence. Therefore, it can be effectively utilized in multi-class classification tasks in text mining [23].

We employ Pipeline class from Scikit-Learn for using CountVectorizer and TF-IDF transformation of the train and test data sets. Therefore, we can convert arbitrary textual data into numerical features to train the selected baseline methods. Also, we convert TF-IDF vectorized dataset to the Attribute-Relation File Format (ARFF) because our Active Learning method is implemented in JCLAL library and this library requires the train and test datasets in ARFF file format.

In the JCLAL library, we perform 10-fold cross-validation as an evaluation method. In this method, it initially shuffles the dataset randomly and split the dataset into 10 groups. For each unique group, it selects the current group as a test dataset and the remaining groups are trained. Then, it fits a model on the training set and evaluates it on the current test set. After that, it keeps the evaluation score for each group and removes the current model. Finally, it uses each evaluation score of 10 samples to summarize the overall performance of the given method.

We define two different stopping criteria to our Pool-based

sampling Active Learning scenario. The first criteria count the number of iterations and it stops until reaching 150 iterations. The second criterion stops while there exists no tweet without an emoji label in the unlabelled dataset. To declare the labelled and unlabelled datasets, we employ the Resample method is chosen from the JCLAL library and set the percentage of the selected parameter to 5, 10 and 20. Therefore, 5% of the generated emoji dataset randomly chosen as the labelled dataset and the rest set to the unlabelled dataset. We define use cases as SOM+Bin-Min-5, SOM+Bin-Min-10, and SOM+Bin-Min-20 which depend on the Sequential Minimal Optimization method using Binary Minimum strategy for the given Resample parameter.

We compare our selected Binary Minimum strategies with three baselines for the multi-class classification purposes as denoted in Table 2. We apply 10-fold cross-validation in which it splits all the samples in 10 groups of samples having equal sizes. Then, we train SOM-based methods using 9 folds and one-fold left out is used for the test. Hence, we compute mean values for precision, recall, F1 and accuracy measures of the 10 training examples. Multinomial Naive Bayes method has the lowest performance among the three baselines and Linear SVM is better than Logistic Regression.

Table 2. Comparison of the proposed method with baselines

Method	Precision	Recall	F1-score	Acc.
Linear SVM	0.891	0.891	0.891	0.912
MLR	0.884	0.886	0.885	0.908
MNB	0.852	0.846	0.849	0.874
SOM+Bin-Min-5	0.895	0.893	0.893	0.916
SOM+Bin-Min-10	0.903	0.901	0.901	0.922
SOM+Bin-Min-20	0.903	0.901	0.902	0.923

The performance of the proposed method outperforms three baselines in all Resample parameter selections. While SOM+Bin-Min-10 and SOM+Bin-Min-20 perform slightly better than SOM+Bin-Min-5, they have almost the same F1-score. Hence, we select the SOM+Bin-Min-10 model as the best method because it requires less labelled dataset than the SOM+Bin-Min-20 model. We also observe the performance of each emoji class for the SOM+Bin-Min-10 model as illustrated in Table 3.

Table 3. Results of the 10 most frequent emojis using SOM+Bin-Min-10

E. Code	P	R	F1	Acc.
U+1F600	0.942	0.891	0.916	0.928
U+1F602	0.918	0.91	0.914	0.919
U+1F60D	0.906	0.878	0.892	0.897
U+1F60E	0.899	0.899	0.899	0.905
U+2764	0.864	0.9	0.882	0.893
U+1F4AA	0.912	0.916	0.914	0.927
U+1F64F	0.897	0.912	0.904	0.915
U+1F615	0.884	0.914	0.899	0.903
U+1F48B	0.902	0.897	0.899	0.908
U+1F620	0.906	0.884	0.895	0.906

Table 3 indicates that the overall performance of classes including more samples than others achieved better results for the top two emojis. However, F1 scores are not a good indicator for the rest of emojis and there exists no dependency between F1 score and the frequency of the emojis.

5. Conclusion

Emojis have been widely used in social media platforms to emphasize the emotions and feelings underlying in the texts. In this study, we provide an emoji prediction method as a multi-class classification task. We trained Naive Bayes, Logistic Regression and Support Vector Machines 3 different baselines on the generated Turkish emoji dataset.

We perform Pool Based Sampling Scenario which includes SMO classifier and Multi-Class BinMin as the query strategy in Active Learning. We select SVM, Multinomial NB and Logistic Regression as three baselines supervised learning methods. The experimental results show that we outperform these baselines and F1-score of the proposed method reaches 0.902 if the Resample parameter is set to 20. Also, we apply three different Resample parameter to the current method to examine the importance of the percentage of the unlabelled dataset in the overall performance. If we choose 10% and 20% of labelled dataset and the rest is coming from the unlabelled dataset, F1 scores are better than 5% selection. But these results are close to each other and it indicates that we can perform 10% selection as the final selected parameter. Because 20% of selection cannot improve the results significantly.

As future studies, we also examine recent vector representations rather than using Count vectorizer with TF-IDF transformation. We examine other Active Learning methods with different strategies and compare them to the current method. In addition, we will increase the total number of emoticons labelled dataset to compare our active learning-based method with recent neural network models.

References

[1] F. Barbieri, F. Ronzano, and H. Saggion, "What does this emoji mean? a vector space skip-gram model for twitter emojis," in Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), 2016, pp. 3967–3972.

[2] P. K. Novak, J. Smailovic, B. Sluban, and I. Mozetic, "Sentiment of emojis," PloS one, vol. 10, no. 12, p. e0144296, 2015.

[3] B. Eisner, T. Rocktaschel, I. Augenstein, M. Bosnjak, and S. Riedel, "emoji2vec: Learning emoji representations from their description," arXiv preprint arXiv:1609.08359, 2016.

[4] B. Settles, "Active learning, "Synthesis Lectures on Artificial Intelligence and Machine Learning, vol. 6, no. 1, pp. 1–114, 2012.

[5] F. Barbieri, M. Ballesteros, and H. Saggion, "Are emojis predictable?" in Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers,2017, pp. 105–111.

[6] X. Li, R. Yan, and M. Zhang, "Joint emoji classification and embedding learning," in Asia-Pacific Web (APWeb) and Web-Age Information Management (WAIM) Joint Conference on Web and Big Data. Springer, 2017, pp. 48–63.

[7] O. Coban, B. Ozyer, and G. T. Ozyer, "Sentiment analysis for turkish twitter feeds," in 2015 23nd Signal Processing and Communications Applications Conference (SIU). IEEE, 2015, pp. 2388–2391.

[8] M. Shiha and S. Ayvaz, "The effects of emoji in sentiment analysis," Int. J. Comput. Electr. Eng. (IJCEE.), vol. 9, no. 1, pp. 360–369, 2017.

[9] C. U. Yurtoz and I. B. Parlak, "Measuring the effects of emojis on Turkish context in sentiment analysis," in2019 7th International Symposium on Digital Forensics and Security (ISDFS). IEEE, 2019, pp. 1–6.

[10] R. Velioglu, T. Yıldız, and S. Yıldırım, "Sentiment analysis using learning approaches over emojis for turkish tweets," in2018 3rd International Conference on Computer Science and Engineering (UBMK). IEEE, 2018, pp. 303–307.

[11] Y.-Y. Yang, S.-C. Lee, Y.-A. Chung, T.-E. Wu, S.-A. Chen, and

H.-T. Lin, “libact: Pool-based active learning in python,” arXiv preprint arXiv:1710.00379, 2017.

[12] O. Reyes, E. Perez, M. Del Carmen Rodriguez-Hernandez, H. M.Fardoun, and S. Ventura, “Jclal: a java framework for active learning,” The Journal of Machine Learning Research, vol. 17, no. 1, pp. 3271–3275, 2016.

[13] G. Tsoumakas, E. Spyromitros-Xioufis, J. Vilcek, and I. Vlahavas, “Mulan: A java library for multi-label learning,” Journal of Machine Learning Research, vol. 12, no. Jul, pp. 2411–2414, 2011.

[14] S. Keerthi, S. Shevade, C. Bhattacharyya, and K. Murthy, “Improve-ments to platt’s smo algorithm for svm classifier design,” Neural Computation, vol. 13, no. 3, pp. 637–649, 2001.

[15] T. Hastie and R. Tibshirani, “Classification by pairwise coupling,” in Advances in Neural Information Processing Systems, M. I. Jordan, M. J.Kearns, and S. A. Solla, Eds., vol. 10. MIT Press, 1998.

[16] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H.Witten, “The weka data mining software: an update,” ACM SIGKDD explorations newsletter, vol. 11, no. 1, pp. 10–18, 2009.

[17] K. Brinker, “On active learning in multi-label classification,” in From Data and Information Analysis to Knowledge Engineering. Springer, 2006, pp. 206–213.

[18] M. del Pilar Salas-Zrate, M. A. Paredes-Valverde, M. ngel Rodriguez-Garca, R. Valencia-Garca, and G. Alor-Hernndez, “Automatic detectionof satire in twitter: A psycholinguistic-based approach,”Knowledge-Based Systems, vol. 128, pp. 20 – 33, 2017. [Online].Available:http://www.sciencedirect.com/science/article/pii/S0950705117301855

[19] J. Cotelo, F. Cruz, J. Troyano, and F. Ortega, “A modular approach for lexical normalization applied to spanish tweets,” Expert Systems with Applications, vol. 42, no. 10, pp. 4743 – 4754, 2015. [Online].Available:http://www.sciencedirect.com/science/article/pii/S0957417415000962

[20] O. Gorgun and O. T. Yildiz, “A novel approach to morphological disambiguation for turkish,” in Computer and Information Sciences II. Springer, 2011, pp. 77–83.

[21] J. Han, J. Pei, and M. Kamber, Data mining: concepts and techniques. Elsevier, 2011.

[22] D. G. Kleinbaum, K. Dietz, M. Gail, M. Klein, and M. Klein, Logistic regression. Springer, 2002.

[23] C. D. Manning, P. Raghavan, and H. Schutze, “Introduction to infor-mation retrieval? cambridge university press 2008,” Ch, vol. 20, pp. 405–416.