

# We are IntechOpen, the world's leading publisher of Open Access books Built by scientists, for scientists

6,900

Open access books available

186,000

International authors and editors

200M

Downloads

Our authors are among the

154

Countries delivered to

TOP 1%

most cited scientists

12.2%

Contributors from top 500 universities



WEB OF SCIENCE™

Selection of our books indexed in the Book Citation Index  
in Web of Science™ Core Collection (BKCI)

Interested in publishing with us?  
Contact [book.department@intechopen.com](mailto:book.department@intechopen.com)

Numbers displayed above are based on latest data collected.  
For more information visit [www.intechopen.com](http://www.intechopen.com)



# Text Classification on the Instagram Caption Using Support Vector Machine

*Setiawan Hadi and Paquita Putri Ramadhani*

## Abstract

Instagram is one of the world's top ten most popular social networks. Instagram is the most popular social networking platform in the United States, India, and Brazil, with over 1 billion monthly active users. Each of these countries has more than 91 million Instagram users. The number of Instagram users shows the various reasons and goals for them to play this social media. Social Media Marketing does not escape being one of the purposes of using Instagram, with benefits to place a market for their products. Using text classification to categorize Instagram captions into organized groups, namely fashion, food & beverage, technology, health & beauty, lifestyle & travel, this paper is expected to help people know the current trends on Instagram. The Support Vector Machine algorithm in this research is used in 66171 post captions to classify trending on Instagram. The TF-IDF (Term Frequency times Inverse Document Frequency) method and percentage variations were used for data separation in this study. This study result indicates that the use of SVM with a percentage ratio 70% of dataset for training and 30% of dataset for testing produces a higher level of accuracy compared to the others.

**Keywords:** Instagram, Support Vector Machine, Text Classification, TFIDF, Social Media

## 1. Introduction

Currently, the internet and humans cannot be separated because of the large amount of information and knowledge available on the internet with its ability to facilitate access to various things. In addition to information disclosure, the internet is also used as a place to share experiences and hobbies through social media [1].

Obtaining an overview of social media, according to Wikipedia, social media is an online platform that allows individuals to easily join, share, social networks, wikis, forums, and create blogs. Blogs, social networks, and wikis are the most common social media used by people worldwide. As of August 2017, Instagram is the sixth most popular social media platform with 700 million members.

This social media platform, commonly called IG or Insta, is an image and video sharing application that facilitates users to upload photos and videos, apply digital

filters to photos and videos, and also share them on other social media [2]. Moreover, Instagram also has several other functions, namely:

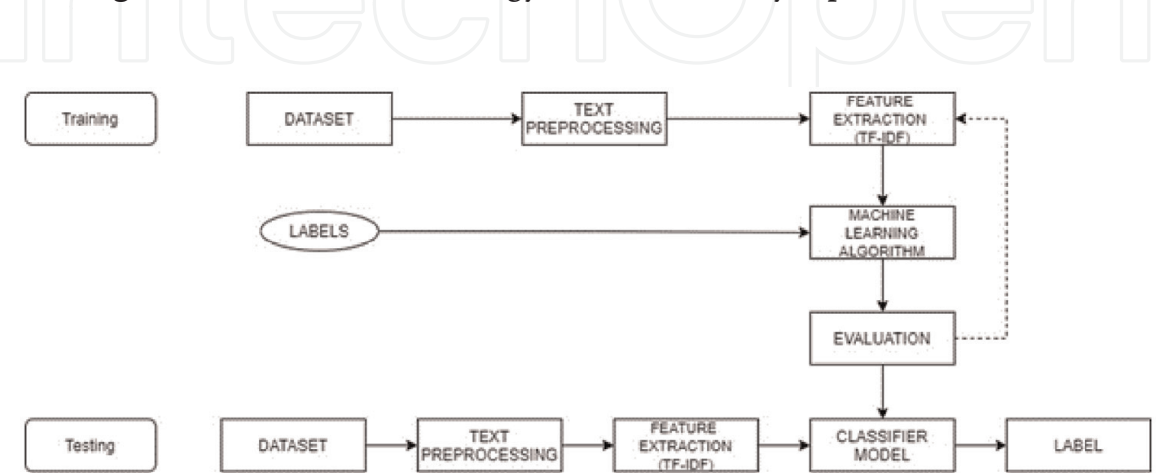
- 1. Interact with fellow Instagram users
- 2. Share recommendations
- 3. Online marketing
- 4. Share hobbies or other interests

At first, social media was just a way for people to communicate with one another. As technology advances, social media allows people to express themselves as creators and thinkers, rather than just as observers. Which activities can be facilely done using Instagram. Due to the increasingly massive use of social media, marketing through social media appears to be the best option in developing their business [3].

The caption in every Instagram post is one way to attract the audience’s interest to buy the goods or services being traded [4]. Audiences can interact with or respond to the post. Observations show that a post gets significantly different interactions, depending on the content of the image and the caption. When an image is uploaded with a specific caption, especially using a hashtag, the post can become a trend. The profile of a person who is a potential target market, or demographic segmentation, behavioral segmentation, and lifestyle segmentation, is related to interests. These things allow marketers to know who is paying attention and interest in the trend. According to Shopify.co.id, there are several trending Instagram categories in 2020, namely Fashion, Food & Beverage, Technology, Health & Beauty, and Lifestyle & Travel.

We can conclude that the classification of Instagram captions plays a significant role in mapping the development of trends on the platform. By knowing the latest people’s favorite trends, new business people have the convenience of promoting their brand. The Instagram posts trend can be known through the text classification method. Is the trend towards Fashion, Food & Beverage, Technology, Health & Beauty, or Lifestyle & Travel? We can find out by using the Support Vector Machine algorithm.

In **Figure 1** below, the methodology used in this study is presented.



**Figure 1.**  
Text classification.

1.1 Data collections

There are two different types of datasets: data training (CSV files) and data crawling JSON as the data testing. 66.171 data are found in the data training, which contains username, caption, and labeling. There are 1.894 Instagram captions obtained for data testing. The caption data retrieval will be processed to produce a certain weight, which will be used later during the Instagram caption data classification process.

The data in **Table 1** is then divided into five categories; Fashion, Food & Beverage, Technology, Health & Beauty, and Lifestyle & Travel. The **Table 2** shows the proportion of the amount of data in each category:

Username	Caption	Labeling
rajvegad055	#viral #top #instatop #public #photography #editz #pose #models #look #attitude #style #bollywood #actorslife #hairstyle	1
Flexkulture	#fashion #fashionista #streetstyle #streetwearfashion #streetwear #hype #hypebeast #highfashion #offwhite #supreme #bape #balenciaga #louisvuitton #gucci #yeezy #lit #fire #drip #trending #trend #trendy #trendsetter #fashionblogger #streetstylefashion #culture #style #sneakers	1
Ishovonn	#insta #instagram #travelgram #travel #instahub #instacool #instagramhub #nature #landscape #landscapephotography #naturephotography #natureporn #ourplanetdaily #photography #fashion #streetphotography #natureaddict #naturelovers #earth #travelphotography #traveling #travelblogger #vsco #vscofilter #vscocam #canonphotographer #rainforest #cairns #australia #worl	1
Lytingz	#friendship #friends #instagood #foodie #foodlover #foodblogger #instagood #ootd #pandor #placetoeatjkt #tea #beverage	2
Willyamyantobong	#bar #band #restaurant #europeanfood #asianfood #culinary #hangout #dinner #night #friends #photooftheday #steik #foodblogger #fashionblogger #ootd #asian #asianboys #asianguys #asianwoman #candle #smile #happytummy #happy #livemusic #drink #saturdaynight #interior #song #travelblogger #weekend	2
feedmelicious	#feedmelicious #sushi #fish #sushiroll #feedme #feedmelicious #wasab #travelfood #sandiego #yummy #food #foodporn #chopsticks #tea #soysauce #tea #eatme	2
erateknologi4.0	#technology #tech #innovation #business #iphone #engineering #programming #science #design #apple #electronics #software #computer #gadgets #instagood #coding #follow #android #love #instatech #technews #geek #developer #startup #programmer #instagram #future #gadget #smartphone #bhfyp	3
tiansetia84	#technology #tech #innovation #business #iphone #engineering #electronics #science #instagood #programming #gadgets #design #art #geek #computer #coding #software #apple #android #love #smartphone #gadget #techie #developer #samsung #instatech #engineer #music #ai #bhfyp	3
ndiie_	#work #working #job #myjob #office #company #bored #grind #mygrind #dayjob #ilovemyjob #dailygrind #photooftheday #business #biz #life #workinglate #computer #instajob #instalife #instagood #instadaily	3
ayunaza69	#black #flower #pink #beauty	4
oneanonly143	#nature #beauty #liveyourlife #love	4

Username	Caption	Labeling
nagachuba_village	#fb #instagram #beauty	4
djricky07	#djricky #lifestyle #motivation #goals #entrepreneur #inspiration #busines #livelife kingsize #nightlife #fame #instagram #instafashion #instapic #photoshoot #photooftheday	5
andreigorlov	#itunes #applemusic #music #electronic #lightstorm #usa #apple #newmusic #new #travel #ipod #beats #epic #welcome #gramtrend #andreigorlov #time #apple #news #spotify #insta #love #andreigorlovofficial #photooftheday #beauty #amazing #pluto #nasa #instagood	5
athayara_	#STEPA #STEPA #STEPA #hut #smp #smpnegeri #stepa #stepamadiun #kotamadiun #madiunkota #kotagadis #instabirthday #photooftheday #LATEPOST	5

**Table 1.**  
*Instagram caption data.*

Labelling	Category/Class	Data
1	Fashion	12,638
2	Food & Beverage	8,338
3	Technology	1,385
4	Health & Beauty	22,816
5	Lifestyle & Travel	20,994

**Table 2.**  
*Proportion of the amount of data In each category/class.*

We can see in the **Table 2** that shows an imbalanced dataset, where a disproportionate ratio is found in each class. This disproportionate ratio can be spotted in the Health & Beauty and Technology category data, which has a significant difference in data. This imbalanced dataset will impact the prediction process in each class later. With the imbalanced dataset, the model will tend to predict the majority class data. Meanwhile, the minority class will be treated as noise or even ignored on some occasions. Due to that, there might be misclassification of the minority class compared to the majority class. In this research, the way to resolve the imbalanced dataset is by using the performance matrix, which is the  $F_1$  score.

1.1.1 Text preprocessing

The text preprocessing step is the beginning part of text mining. In text mining, preprocessing is the act of transforming poorly formatted input into structured data that meets the demands of the process.

The preprocessing stage is presented in **Figure 2**. After collecting the data, the next process was text processing. It included case folding, tokenizing, and cleaning.

**Case folding** is the process of converting the letters contained in the text into lowercase letters. Characters other than letters in the A-Z alphabet are omitted. This process was carried out due to the inconsistent use of lowercase and uppercase letters in Instagram captions. Case Folding aims to convert all data in the form of Instagram



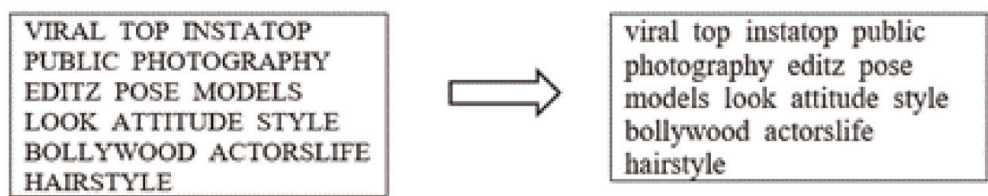


**Figure 2.**  
*Text preprocessing.*

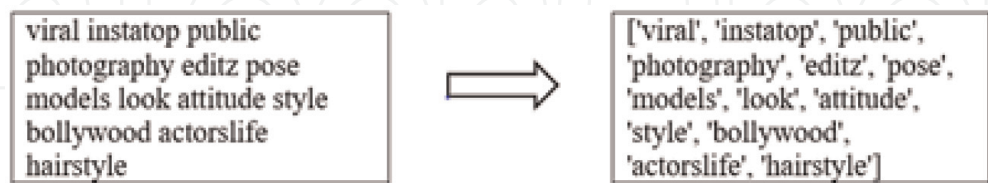
captions to conform to the standard, which usually uses lowercase letters [5]. The other characters which are not letters or numbers, like punctuation and space, will be considered as delimiter. The other characters which are not letters or numbers, like punctuation and space, will be considered as delimiter. The illustration is displayed in **Figure 3**.

**Tokenizing** is a process to divides a large number of characters in a text into a single word unit by distinguishing particular characters required as a word separator [5]. Each word is identified or separated with another using space character, so this tokenizing process relies on space characters in the document to separate the words. The process is illustrated in **Figure 4**.

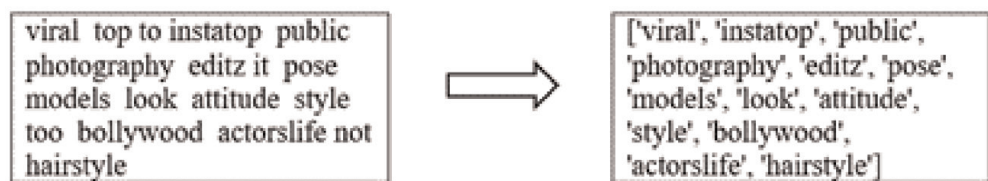
**Filtering** is a method that uses a stoplist (removing unnecessary words) or wordlist to extract certain key words from the token results (including crucial words). Some English stopword examples are “the” “from”, “and”, and others. The meaning behind the stopword use is to remove words with low information in a text to focus on the essential words to replace them. Filtering is done by determining what terms will be used to represent a document, where a document describes each of its contents and differs from one another. This process is illuatrated in **Figure 5**.



**Figure 3.**  
*Case folding process.*



**Figure 4.**  
*Tokenizing process.*



**Figure 5.**  
*Filtering (Stopword removal) process.*

## 2. TF-IDF

The next step after text processing is TF — IDF method. At this stage, each word is assigned a weight based on how frequently it appears in the manuscript or document [6]. The computation of Term Frequency (TF) and Inverse Document Frequency (IDF) is also included in this technique (IDF). The steps are as follows:

1. Term Frequency (TF).
2. Inverse Document Frequency (IDF)
3. Term Frequency-Inverse Document Frequency (TF — IDF)

**TF (Term Frequency)** means the number of occurrences or the frequency of words in a document is calculated. The larger the conformity value, the more frequently a phrase appears in a text, indicating that it has a high TF [2]. Here is the formula from the TF:

$$TF = \begin{cases} 1 + \log_{10}(f_{t,d}) & , f_{t,d} > 0 \\ 0 & , f_{t,d} = 0 \end{cases} \quad (1)$$

A frequency term ( $t$ ) in a document ( $d$ ) is the value of  $f_{t,d}$ .

**IDF (Inverse Document Frequency)** means the distribution of a term in a collection of related texts is calculated. The relationship between the terms available in the text is also shown through this [7]. The less text a particular term contains, the larger the IDF. Here is the formula from the IDF:

$$IDF_t = \log \left( \frac{N}{df_t} \right) \quad (2)$$

$N$ : The number of text documents  $df_t$ : The total number of documents that containing the phrase “t-word” (according to the referred term).

**TF — IDF** is the multiplication of the results of the weighting of the frequency of a term and the frequency of the document inversely related to that term [7]. Here is the formula from the TF — IDF:

$$w_{ij} = TF \times IDF_t \quad (3)$$

$TF$ : Term Frequency  $IDF_t$ : Inverse Document Frequency.

## 3. Support vector machine (SVM) model

A classification model named Support Vector Machine (SVM) is used in this method. Support Vector Machine is a supervised learning model. In its application, several linear functions of high dimensional space (feature space) are utilized. This linear function aims to find the best hyperplane in maximizing each class gap [8].

In short, support vector machine is a linear classifier. However, in some nonlinear problems, this model can also be used with some improvements [9], which are needed because not all data is linearly divided. This results in non-optimal results if linear SVM is still applied.

The radial basis function (RBF) kernel was used to change the SVM modeling process from linear to non-linear [10]. Generally, the RBF kernel is used for all types of data as a linear data separator. The RBF kernel has two parameters, namely Gamma and Cost.

The Cost parameter is used for SVM optimization so that misclassification in the training dataset sample nghwaes not occur. Meanwhile, to measure the influence given by each training dataset sample, the Gamma parameter is used [11]. A low or high value is indicated by the use of this parameter. Low or high values are described as “far” and “near”. The formula below is the RBF Kernel equation:

$$K_{(x,z)} = \exp \left[ -\gamma \|x - z\|^2 \right] \quad (4)$$

#### 4. $F_1$ score

The value of accuracy in testing the data is known by the  $F_1$  score, which is the average of Precision and Recall, where both metrics are calculated simultaneously [10]. Precision describes the degree of precision between the required data and the model's predicted outputs [10]. The percentage of success of a model in recovering information is represented through Recall. The formula for the  $F_1$  score is as follows:

$$F_1 \text{ Score} = 2 * \frac{\text{precision} * \text{recall}}{\text{precision} + \text{recall}} \quad (5)$$

The  $F_1$  score calculation can be used as an evaluation standard from the predictive classification result if there is a class imbalance in the data.

The following are the steps taken to conduct this research:

1. The data from this study are classified into several types. Each type is labeled as follows; Fashion, Food & Beverage, Technology, Health & Beauty, and Lifestyle & Travel.
2. Case-folding and tokenizing are carried out in the data processing process by applying them to any text used in data training or data testing. After that, a new document is obtained in order to proceed to the following step.
3. At the feature extraction stage, TF-IDF is implemented in data training or data testing.
4. In the data split stage, ratios of 70:30, 60:40, 50:50, and 40:60 were used for data training.
5. This series ends with an evaluation stage in which the  $F_1$  score is used to determine the prediction results on the training data.



4.1 Result and discussion

This research begins by analyzing the dataset that has been prepared to determine whether the data has missing values, data imbalances, and other problems in the data. Proceed to the preprocessing stage to remove symbols, emoji, number punctuations, and white/multiple spaces in the text. Filtering is also done to remove words that are stop words. Then, To identify the frequency of occurrence of a word in the document, the data is transformed into vector form, and the value of Term Frequency (TF) and Inverse Document Frequency (IDF) for each token (word) is calculated. The clean and weighted data is then divided into train and testing groups with varying ratios. The support vector machine method and the radial basis function (RBF) kernel are used to classify Instagram caption data. The whole process ends with evaluating the algorithm performance using the  $F_1$  score to overcome the imbalanced data. The difference in the distribution of train data and testing data proportion aims to see whether there is an effect of the training data and testing data proportion on the results of the  $F_1$  score.

The outcome of the analysis is as follows:

From **Table 3**, we can see the  $F_1$  score is obtained from the experiment. The  $F_1$  score is generated using a distinct proportion of training and testing data and the results of Recall value and Precision. The results show that a bigger proportion of data training, compared to the data testing, will produce a more significant  $F_1$  score compared to the other proportions.

**Tables 4–7** show particular findings for Precision value, Recall, and  $F_1$  Score in each category with a varied proportion of data training and testing (70:30, 60:40, 50:50, and 40:60, respectively). The following is the result of the calculation for Precision value, Recall, and  $F_1$  Score in each data proportion:

In **Table 3** the classification results are presented using the Support Vector Machine algorithm. The average  $F_1$  score is above 88% and the largest  $F_1$  score is the proportion of training and testing data with a proportion of 70:30. These results are obtained through the Kernel Radial Basis Function (RBF). This proves that a larger amount of training data in a model can produce better results. The  $F_1$  scores from each category with different training data share and testing data proportions are shown in **Tables 4–7**. The proportion of data share from training data and testing data generated is 70:30. These results are better, especially in the Technology category.

It might be interesting to split training data set and testing data set with the ratio of 80 per cent training set and 20 per cent test set and perform another experiment using that ratio. The result could give higher or lower accuracy compared with previous experiment. However, based on the references, it will be depends on the method and algorithm used.

Proportion	Precision	Recall	$F_1$ Score
70:30	0.90	0.87	0.8895
60:40	0.90	0.86	0.8876
50:50	0.89	0.85	0.8865
40:60	0.89	0.84	0.8832

**Table 3.**  
*Comparison of precision, recall, and  $F_1$  score for each training and testing proportion.*

	Precision	Recall	$F_1$ score
Fashion	0.83	0.82	0.83
Food & Beverage	0.92	0.87	0.90
Health & Beauty	0.89	0.95	0.92
Lifestyle & Travel	0.90	0.87	0.89
Technology	0.92	0.81	0.86

**Table 4.**  
Proportion of data 70:30 for comparison of precision, recall,  $F_1$  score.

Category	Precision	Recall	$F_1$ score
Fashion	0.83	0.82	0.83
Food & Beverage	0.92	0.86	0.89
Health & Beauty	0.90	0.95	0.92
Lifestyle & Travel	0.89	0.88	0.89
Technology	0.93	0.79	0.85

**Table 5.**  
Proportion of 60:40 for comparison of precision, recall,  $F_1$  score.

Category	Precision	Recall	$F_1$ score
Fashion	0.83	0.82	0.83
Food & Beverage	0.92	0.86	0.89
Health & Beauty	0.90	0.95	0.92
Lifestyle & Travel	0.89	0.88	0.89
Technology	0.93	0.74	0.83

**Table 6.**  
Proportion of 50:50 for comparison of precision, recall,  $F_1$  score.

Category	Precision	Recall	$F_1$ score
Fashion	0.83	0.82	0.82
Food & Beverage	0.92	0.85	0.88
Health & Beauty	0.90	0.94	0.92
Lifestyle & Travel	0.89	0.88	0.88
Technology	0.92	0.71	0.80

**Table 7.**  
Proportion of 40:60 for comparison of precision, recall,  $F_1$  score.

## 5. Conclusion

The conclusions that can be drawn from this research are as follows:

1. In this study, a very good  $F_1$  score, above 88%, was obtained using the Support Vector Machine (SVM) with Kernel Radial Basis Function (RBF).
2. The performance of the SVM algorithm has increased with the use of TF-IDF as a feature extraction method. The possibility of a different reaction from the algorithm, namely by not getting the expected result, can occur if there is untrained data in the data set. Data that has not been validated by experts is untrained data. Sometimes, inaccuracies can result from improper labeling of the source.
3. Model performance may be improved by dividing the data into several proportions. The use of more training also makes it possible to get better model results. This affects the researchers' use of as much data as possible to train the model.


### Author details

Setiawan Hadi\*<sup>†</sup> and Paquita Putri Ramadhani<sup>†</sup>  
Universitas Padjadjaran, Jatinangor, Indonesia

\*Address all correspondence to: setiawanhadi@unpad.ac.id

<sup>†</sup> These authors contributed equally.

### IntechOpen

© 2021 The Author(s). Licensee IntechOpen. This chapter is distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/3.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. 

## References

- [1] Indika D R, Jovita C: Instagram social media as a promotional tool for improving consumer buying interest. *Journal of Applied Business, Polytechnic Ubaya*. 2017; 1: 25–32.
- [2] Chen H: College-Aged Young Consumers' Perceptions of Social Media Marketing: The Story of Instagram. *Journal of Current Issues & Research in Advertising*. 2017; 39: 1–15.
- [3] Ting H, Ming W W P, Run E C D, Choo S L Y: Beliefs about the Use of Instagram: An Exploratory Study. *International Journal of Business and Innovation*. 2015: 2:15–31
- [4] Adegbola O, Gearhart S., Skarda-Mitchell J.: Using Instagram to Engage with (Potential) Consumers: A Study of Forbes Most Valuable Brands' Use of Instagram *The Journal of Social Media in Society*. 2018: 7(2): 232–251
- [5] Sebastiani F: Machine Learning in Automated Text Categorization. *ACM Computing Surveys*. 2001: 34:1-47
- [6] Kulkarni A, Shivananda A : *Natural Language Processing Recipes: Unlocking Text Data with Machine Learning and Deep Learning using Python* 1st ed. Edition. Apress. 2019
- [7] Kedia A, Rasu M: *Hands-On Python Natural Language Processing*. Packt Publishing. 2020
- [8] Géron A: *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*. O'Reilly Media. 2019
- [9] Steinwart I, Christmann A. *Support Vector Machines*. Information Science and Statistics Springer-Verlag; 2008
- [10] Sokolova M, Japkowicz N, Szpakowicz S: Beyond Accuracy, F-Score and ROC: A Family of Discriminant Measures for Performance Evaluation AI. *Advances in Artificial Intelligence Lecture Notes in Computer Science* 4304. 2006: 1015-102
- [11] Schlkopf B, Smola A. J.: *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. The MIT Press. 2001