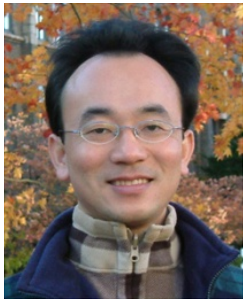
文字探勘
(Text Mining)
情感分析
(Sentiment Analysis)

1082TM10
MBA, BDABI, TKU (E3611) (8480) (Spring 2020)
Mon, 7, 8, 9 (14:10-17:00) (B206)

Chichang Jou
周清江
Associate Professor
副教授
cjou@mail.tku.edu.tw



Min-Yuh Day
戴敏育
Associate Professor
副教授
myday@mail.tku.edu.tw

Dept. of Information Management, Tamkang University
淡江大學 資訊管理學系

2020-05-18

課程大綱 (Syllabus)

週次 (Week)	日期 (Date)	內容 (Subject/Topics)
1	2020/03/02	文字探勘課程介紹 (Course Orientation on Text Mining)
2	2020/03/09	文字探勘基礎：自然語言處理 (Foundations of Text Mining: Natural Language Processing; NLP)
3	2020/03/16	Python自然語言處理 (Python for Natural Language Processing)
4	2020/03/23	處理和理解文本 (Processing and Understanding Text)
5	2020/03/30	文本表達特徵工程 (Feature Engineering for Text Representation)
6	2020/04/06	人工智慧文本分析個案研究 I (Case Study on Artificial Intelligence for Text Analytics I)

課程大綱 (Syllabus)

週次 (Week)	日期 (Date)	內容 (Subject/Topics)
7	2020/04/13	文本分類 (Text Classification)
8	2020/04/20	文本摘要和主題模型 (Text Summarization and Topic Models)
9	2020/04/27	期中報告 (Midterm Project Report)
10	2020/05/04	文本相似度和分群 (Text Similarity and Clustering)
11	2020/05/11	語意分析和命名實體識別 (Semantic Analysis and Named Entity Recognition; NER)
12	2020/05/18	情感分析 (Sentiment Analysis)

課程大綱 (Syllabus)

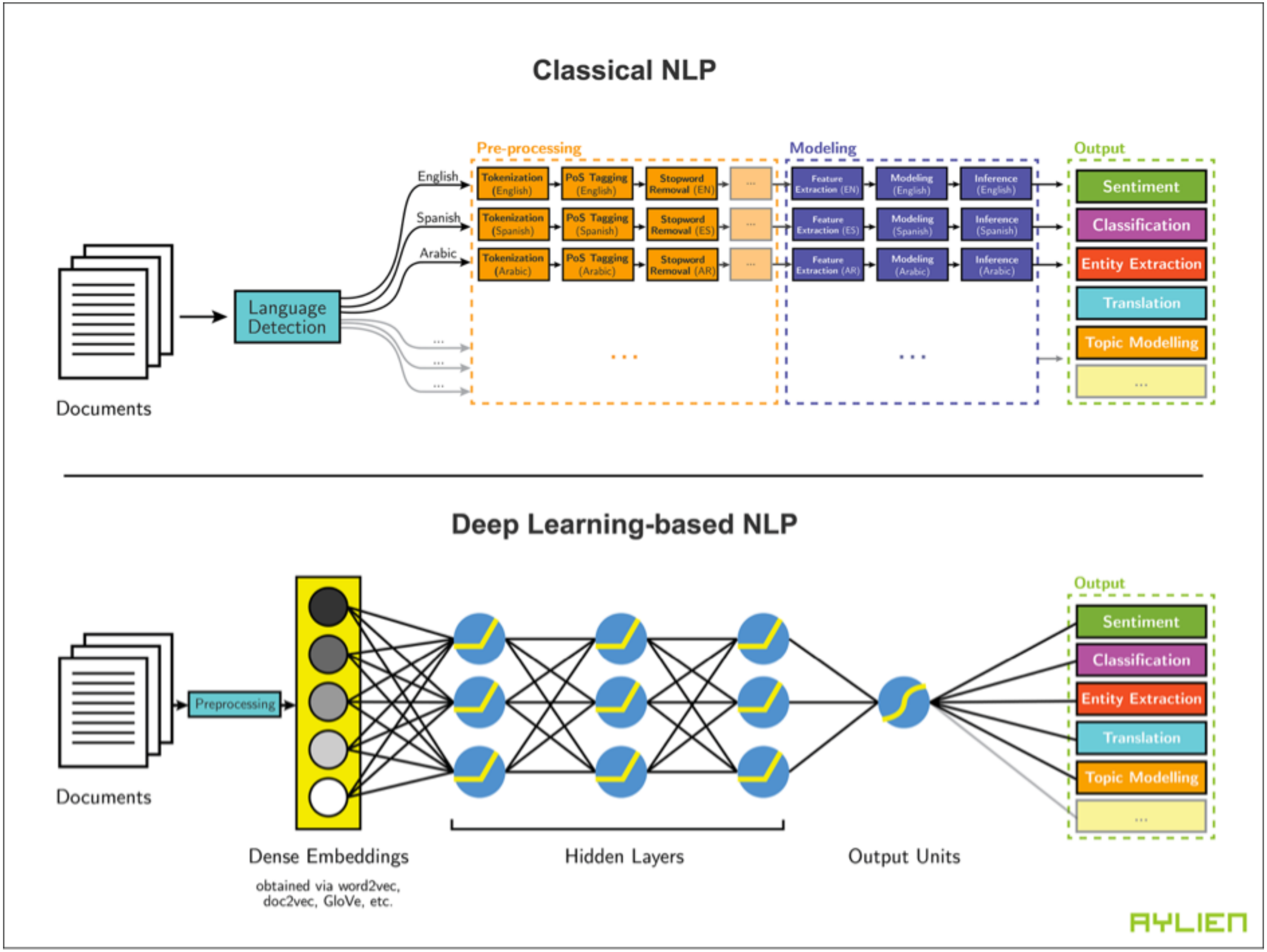
週次 (Week)	日期 (Date)	內容 (Subject/Topics)
13	2020/05/25	人工智慧文本分析個案研究 II (Case Study on Artificial Intelligence for Text Analytics II)
14	2020/06/01	深度學習和通用句子嵌入模型 (Deep Learning and Universal Sentence-Embedding Models)
15	2020/06/08	問答系統與對話系統 (Question Answering and Dialogue Systems)
16	2020/06/15	期末報告 I (Final Project Presentation I)
17	2020/06/22	期末報告 II (Final Project Presentation II)
18	2020/06/29	教師彈性補充教學

Sentiment Analysis

Outline

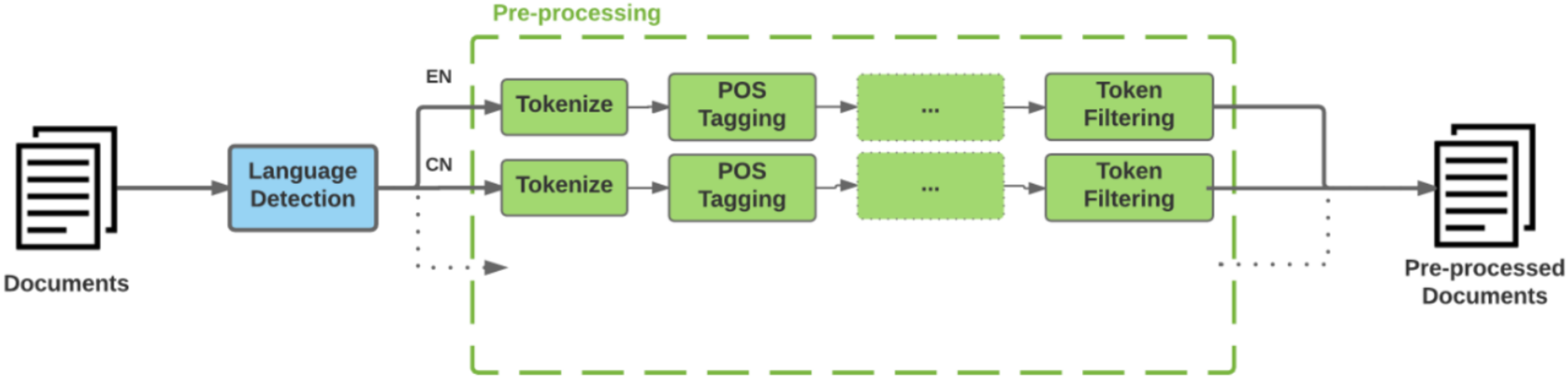
- Unsupervised lexicon-based models
- Traditional supervised machine learning models
- Supervised deep learning models
- Advanced supervised deep learning models

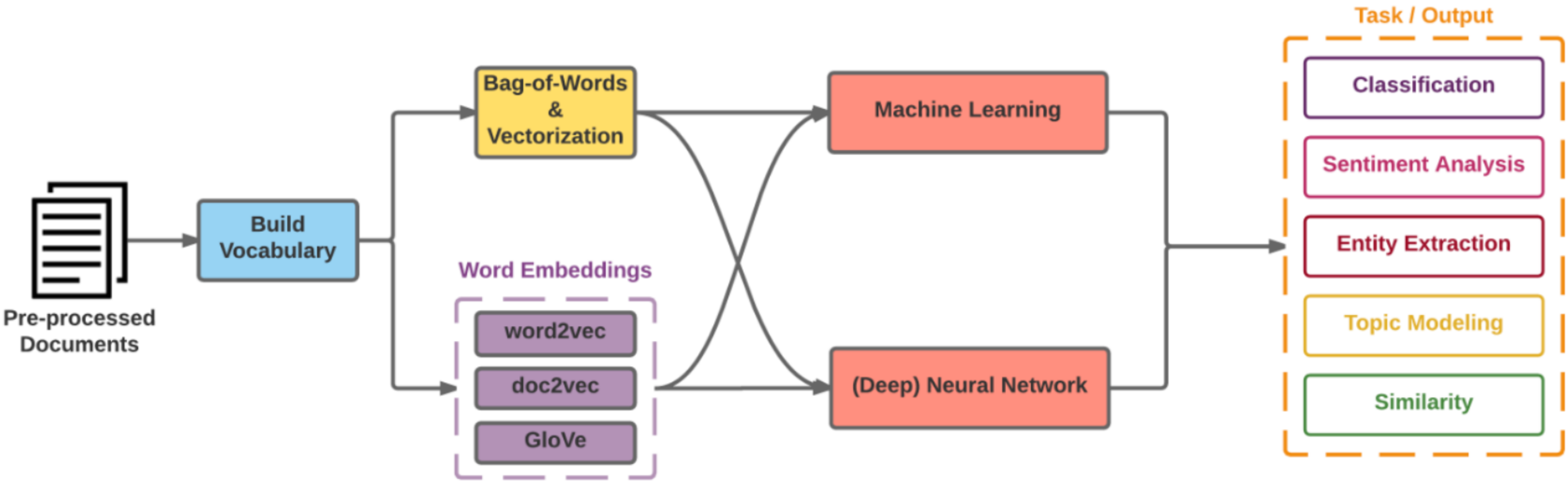
NLP



Source: http://blog.aylien.com/leveraging-deep-learning-for-multilingual/

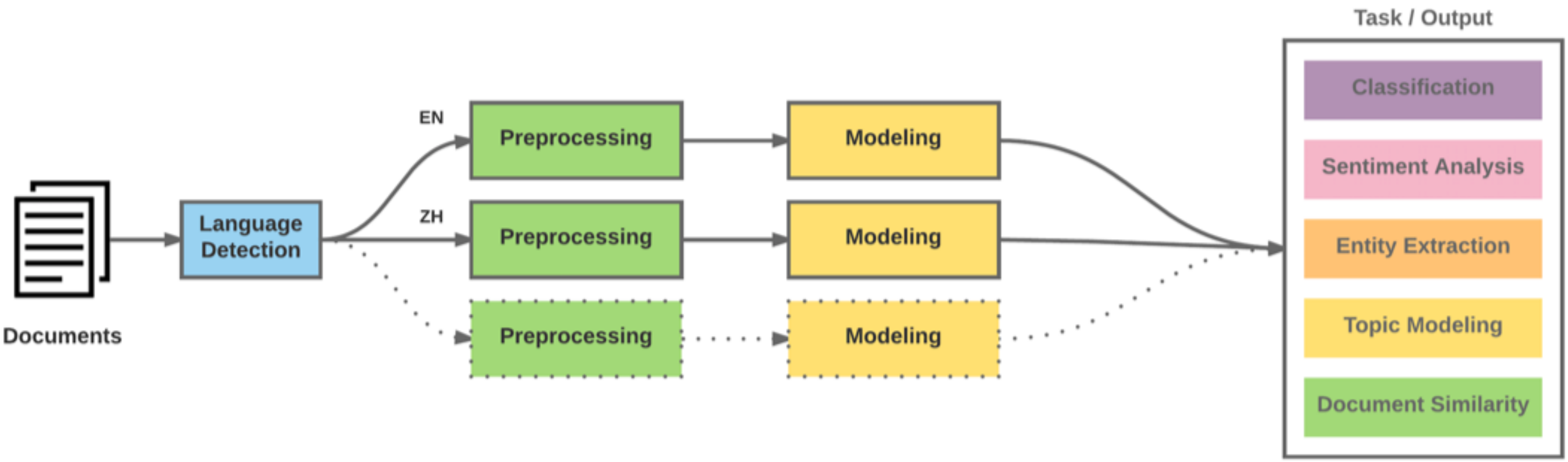
Modern NLP Pipeline





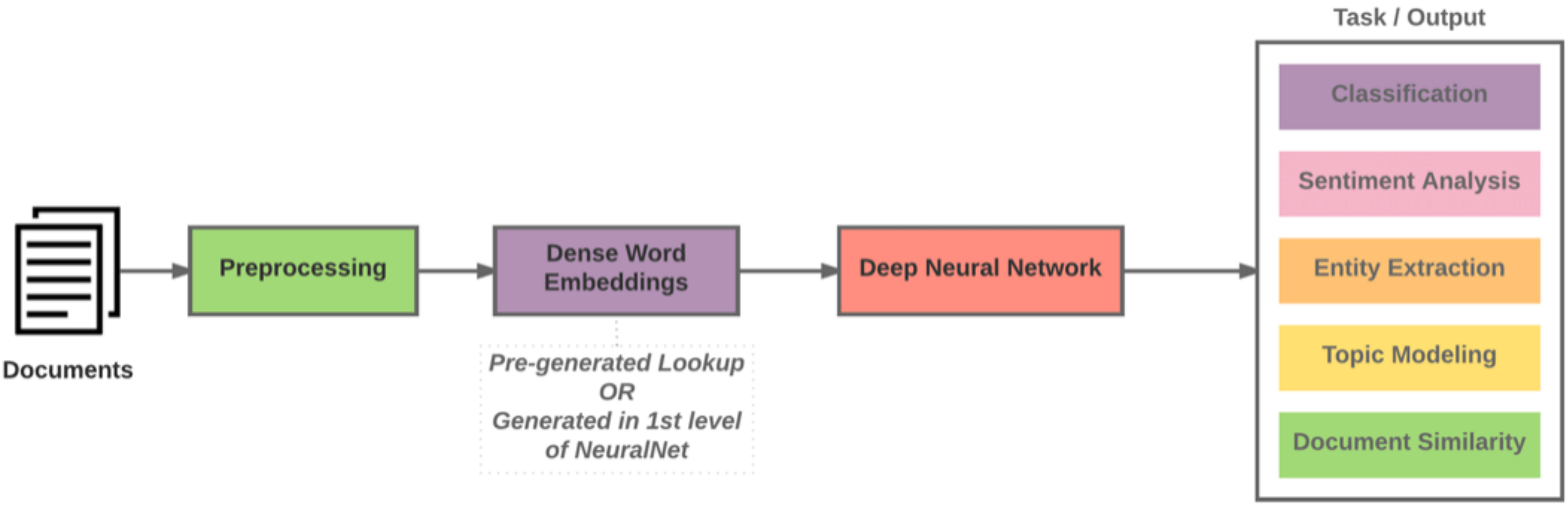
Source: https://github.com/fortiema/talks/blob/master/opendata2016sh/pragmatic-nlp-opendata2016sh.pdf

Modern NLP Pipeline



Source: http://mattfortier.me/2017/01/31/nlp-intro-pt-1-overview/

Deep Learning NLP



Source: http://mattfortier.me/2017/01/31/nlp-intro-pt-1-overview/

Natural Language Processing (NLP) and Text Mining

- Raw text
- Sentence Segmentation
- Tokenization
- Part-of-Speech (POS)
- Stop word removal
- Stemming / Lemmatization
- Dependency Parser
- String Metrics & Matching

word’s stem	word’s lemma
am → am	am → be
having → hav	having → have

Source: Nitin Hardeniya (2015), NLTK Essentials, Packt Publishing; Florian Leitner (2015), Text mining - from Bayes rule to dependency parsing

Large Movie Review Dataset

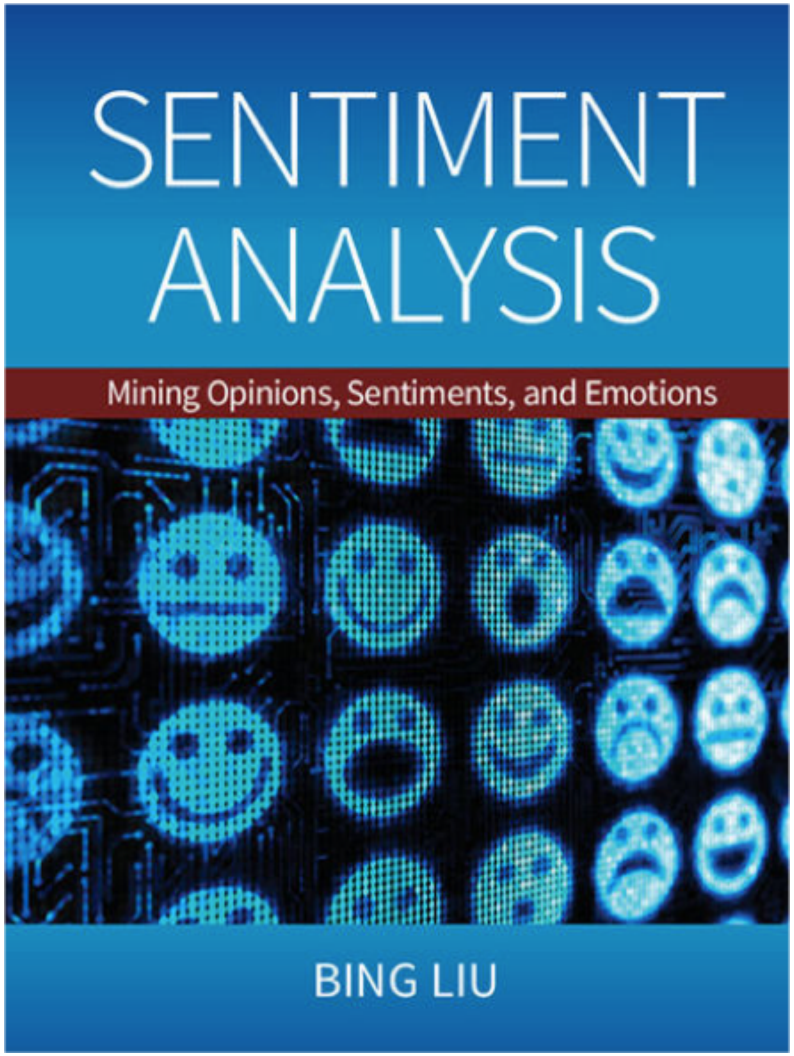
- Large Movie Review Dataset v1.0
  - Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. (2011). Learning Word Vectors for Sentiment Analysis. The 49th Annual Meeting of the Association for Computational Linguistics (ACL 2011).
  - http://ai.stanford.edu/~amaas/data/sentiment/
  - http://ai.stanford.edu/~amaas/data/sentiment/aclImdb_v1.tar.gz

Sentiment Analysis:
Unsupervised Lexicon-Based Models

- Bing Liu’s lexicon
- TextBlob lexicon
- SentiWordNet lexicon
- VADER lexicon
- MPQA subjectivity lexicon
- Pattern lexicon
- AFINN lexicon

Source: Dipanjan Sarkar (2019), Text Analytics with Python: A Practitioner’s Guide to Natural Language Processing, Second Edition. APress.

Bing Liu (2015), Sentiment Analysis: Mining Opinions, Sentiments, and Emotions, Cambridge University Press



http://www.amazon.com/Sentiment-Analysis-Opinions-Sentiments-Emotions/dp/1107017890

Emotions

☺ ☹

Love

Joy

Surprise

Anger

Sadness

Fear

Source: Bing Liu (2011) , “Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data,” Springer, 2nd Edition,

Example of Opinion: review segment on iPhone

"I bought an iPhone a few days ago.

It was such a nice phone.

The touch screen was really cool.

The voice quality was clear too.

However, my mother was mad with me as I did not tell her before I bought it.

She also thought the phone was too expensive, and wanted me to return it to the shop. … "

Source: Bing Liu (2011) , "Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data," Springer, 2nd Edition,

Example of Opinion: review segment on iPhone

“(1) I bought an iPhone a few days ago.

(2) It was such a **nice** phone.

(3) The touch screen was really **cool**.

(4) The voice quality was **clear** too.

☺ +Positive Opinion

(5) However, my mother was mad with me as I did not tell her before I bought it.

(6) She also thought the phone was too **expensive**, and wanted me to return it to the shop. … ”

☹ -Negative Opinion

Source: Bing Liu (2011) , “Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data,” Springer, 2nd Edition,

Sentiment Analysis and Opinion Mining

- Computational study of
  opinions,
  sentiments,
  subjectivity,
  evaluations,
  attitudes,
  appraisal,
  affects,
  views,
  emotions,
  ets., expressed in text.
  - Reviews, blogs, discussions, news, comments, feedback, or any other documents

Source: Bing Liu (2011) , “Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data,” Springer, 2nd Edition,

Research Area of Opinion Mining

- Many names and tasks with difference objective and models
  - Sentiment analysis
  - Opinion mining
  - Sentiment mining
  - Subjectivity analysis
  - Affect analysis
  - Emotion detection
  - Opinion spam detection

Source: Bing Liu (2011) , “Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data,” Springer, 2nd Edition,

Sentiment Analysis

- Sentiment
  - A thought, view, or attitude, especially one based mainly on emotion instead of reason
- Sentiment Analysis
  - opinion mining
  - use of natural language processing (NLP) and computational techniques to automate the extraction or classification of sentiment from typically unstructured text

Applications of Sentiment Analysis

- Consumer information
  - Product reviews
- Marketing
  - Consumer attitudes
  - Trends
- Politics
  - Politicians want to know voters’ views
  - Voters want to know policitians’ stances and who else supports them
- Social
  - Find like-minded individuals or communities

Sentiment detection

- How to interpret features for sentiment detection?
  - Bag of words (IR)
  - Annotated lexicons (WordNet, SentiWordNet)
  - Syntactic patterns
- Which features to use?
  - Words (unigrams)
  - Phrases/n-grams
  - Sentences

Problem statement of Opinion Mining

- Two aspects of abstraction
  - Opinion definition
    - What is an opinion?
    - What is the structured definition of opinion?
  - Opinion summarization
    - Opinion are subjective
      - An opinion from a single person (unless a VIP) is often not sufficient for action
    - We need opinions from many people, and thus opinion summarization.

Source: Bing Liu (2011) , "Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data," Springer, 2nd Edition,

What is an opinion?

- Id: **Abc123** on **5-1-2008** "*I bought an iPhone a few days ago. It is such a nice phone. The touch screen is really cool. The voice quality is clear too. It is much better than my old Blackberry, which was a terrible phone and so difficult to type with its tiny keys. However, my mother was mad with me as I did not tell her before I bought the phone. She also thought the phone was too expensive, …*"
- One can look at this review/blog at the
  - Document level
    - Is this review + or -?
  - Sentence level
    - Is each sentence + or -?
  - Entity and feature/aspect level

Source: Bing Liu (2011) , "Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data," Springer, 2nd Edition,

Entity and aspect/feature level

- Id: **Abc123** on **5-1-2008** "*I bought an iPhone a few days ago. It is such a nice phone. The touch screen is really cool. The voice quality is clear too. It is much better than my old Blackberry, which was a terrible phone and so difficult to type with its tiny keys. However, my mother was mad with me as I did not tell her before I bought the phone. She also thought the phone was too expensive, …*"
- What do we see?
  - Opinion targets: entities and their features/aspects
  - Sentiments: positive and negative
  - Opinion holders: persons who hold the opinions
  - Time: when opinion are expressed

Source: Bing Liu (2011) , "Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data," Springer, 2nd Edition,

Two main types of opinions

- Regular opinions: Sentiment/Opinion expressions on some target entities
  - Direct opinions: sentiment expressions on one object:
    - “The touch screen is really cool.”
    - “The picture quality of this camera is great”
  - Indirect opinions: comparisons, relations expressing similarities or differences (objective or subjective) of more than one object
    - “phone X is cheaper than phone Y.” (objective)
    - “phone X is better than phone Y.” (subjective)
- Comparative opinions: comparisons of more than one entity.
  - “iPhone is better than Blackberry.”

Source: Bing Liu (2011) , “Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data,” Springer, 2nd Edition,

Subjective and Objective

- Objective
  - An objective sentence expresses some factual information about the world.
  - “I returned the phone yesterday.”
  - Objective sentences can implicitly indicate opinions
    - “The earphone broke in two days.”
- Subjective
  - A subjective sentence expresses some personal feelings or beliefs.
  - “The voice on my phone was not so clear”
  - Not every subjective sentence contains an opinion
    - “I wanted a phone with good voice quality”
- ➔ Subjective analysis

Source: Bing Liu (2011) , “Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data,” Springer, 2nd Edition,

Sentiment Analysis
vs.
Subjectivity Analysis

Sentiment Analysis	Subjectivity Analysis
Positive	Subjective
Negative	
Neutral	Objective

A (regular) opinion

- Opinion (a restricted definition)
  - An opinion (regular opinion) is simply a positive or negative sentiment, view, attitude, emotion, or appraisal about an entity or an aspect of the entity from an opinion holder.
- Sentiment orientation of an opinion
  - Positive, negative, or neutral (no opinion)
  - Also called:
    - Opinion orientation
    - Semantic orientation
    - Sentiment polarity

Source: Bing Liu (2011) , “Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data,” Springer, 2nd Edition,

Entity and aspect

- Definition of Entity:
  - An *entity e* is a product, person, event, organization, or topic.
  - e is represented as
    - A hierarchy of components, sub-components.
    - Each node represents a components and is associated with a set of attributes of the components
- An opinion can be expressed on any node or attribute of the node
- Aspects(features)
  - represent both components and attribute

Source: Bing Liu (2011) , "Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data," Springer, 2nd Edition,

Opinion Definition

- An opinion is a quintuple
  (e_j, a_{jk}, so_{ijkl}, h_i, t_l)
  where
  - e_j is a target entity.
  - a_{jk} is an aspect/feature of the entity e_j .
  - so_{ijkl} is the sentiment value of the opinion from the opinion holder on feature of entity at time.
    so_{ijkl} is +ve, -ve, or neu, or more granular ratings
  - h_i is an opinion holder.
  - t_l is the time when the opinion is expressed.
- (e_j, a_{jk}) is also called opinion target

Source: Bing Liu (2011) , “Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data,” Springer, 2nd Edition,

Terminologies

- Entity: object
- Aspect: feature, attribute, facet
- Opinion holder: opinion source

- Topic: entity, aspect

- Product features, political issues

Source: Bing Liu (2011) , "Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data," Springer, 2nd Edition,

Subjectivity and Emotion

- Sentence subjectivity
  - An objective sentence presents some factual information, while a subjective sentence expresses some personal feelings, views, emotions, or beliefs.
- Emotion
  - Emotions are people's subjective feelings and thoughts.

Source: Bing Liu (2011) , "Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data," Springer, 2nd Edition,

Classification Based on Supervised Learning

- Sentiment classification
  - Supervised learning Problem
  - Three classes
    - *Positive*
    - *Negative*
    - *Neutral*

Source: Bing Liu (2011) , “Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data,” Springer, 2nd Edition,

Opinion words in Sentiment classification

- topic-based classification
  - topic-related words are important
    - e.g., *politics, sciences, sports*
- Sentiment classification
  - topic-related words are unimportant
  - **opinion words** (also called **sentiment words)**
    - **that indicate positive or negative opinions** are important, e.g., *great, excellent, amazing, horrible, bad, worst*

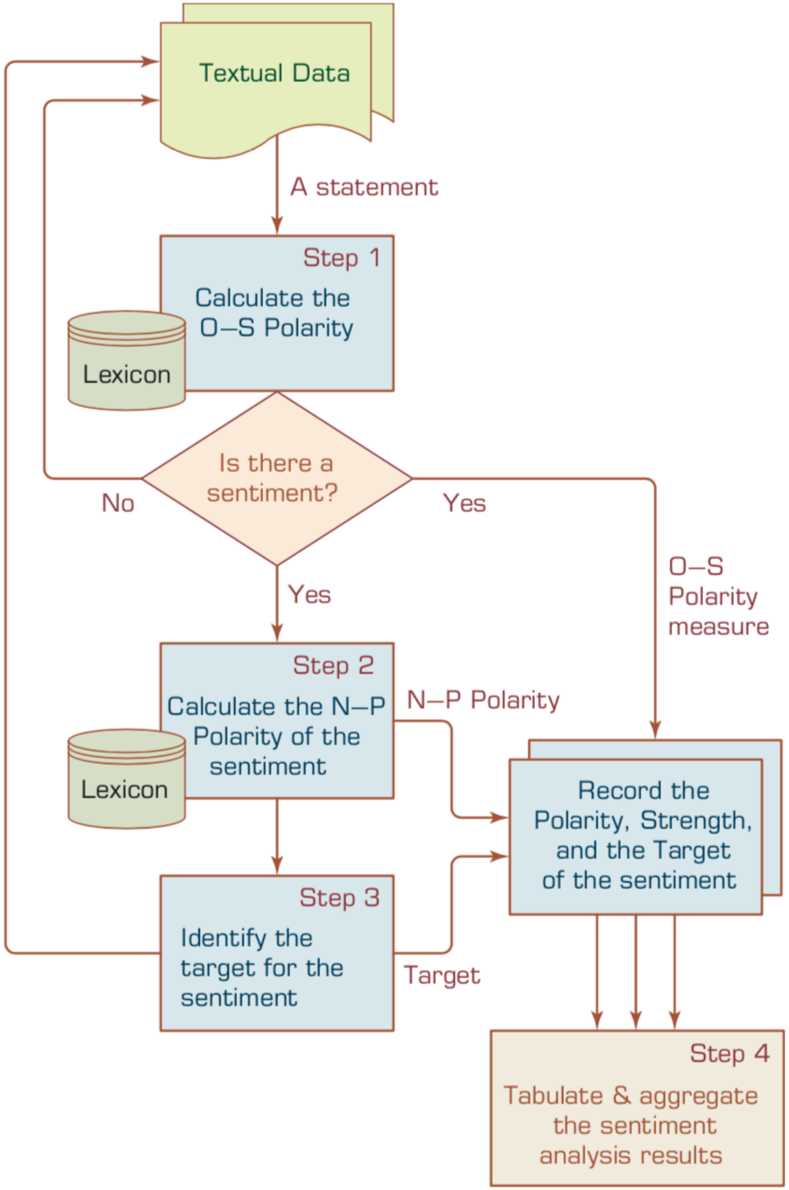
Source: Bing Liu (2011) , “Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data,” Springer, 2nd Edition,

Features in Opinion Mining

- *Terms and their frequency*
  - *TF-IDF*
- *Part of speech (POS)*
  - Adjectives
- *Opinion words and phrases*
  - *beautiful, wonderful, good, and amazing are positive opinion words*
  - *bad, poor, and terrible are negative opinion words.*
  - opinion phrases and idioms,
    e.g., *cost someone an arm and a leg*
- *Rules of opinions*
- *Negations*
- *Syntactic dependency*

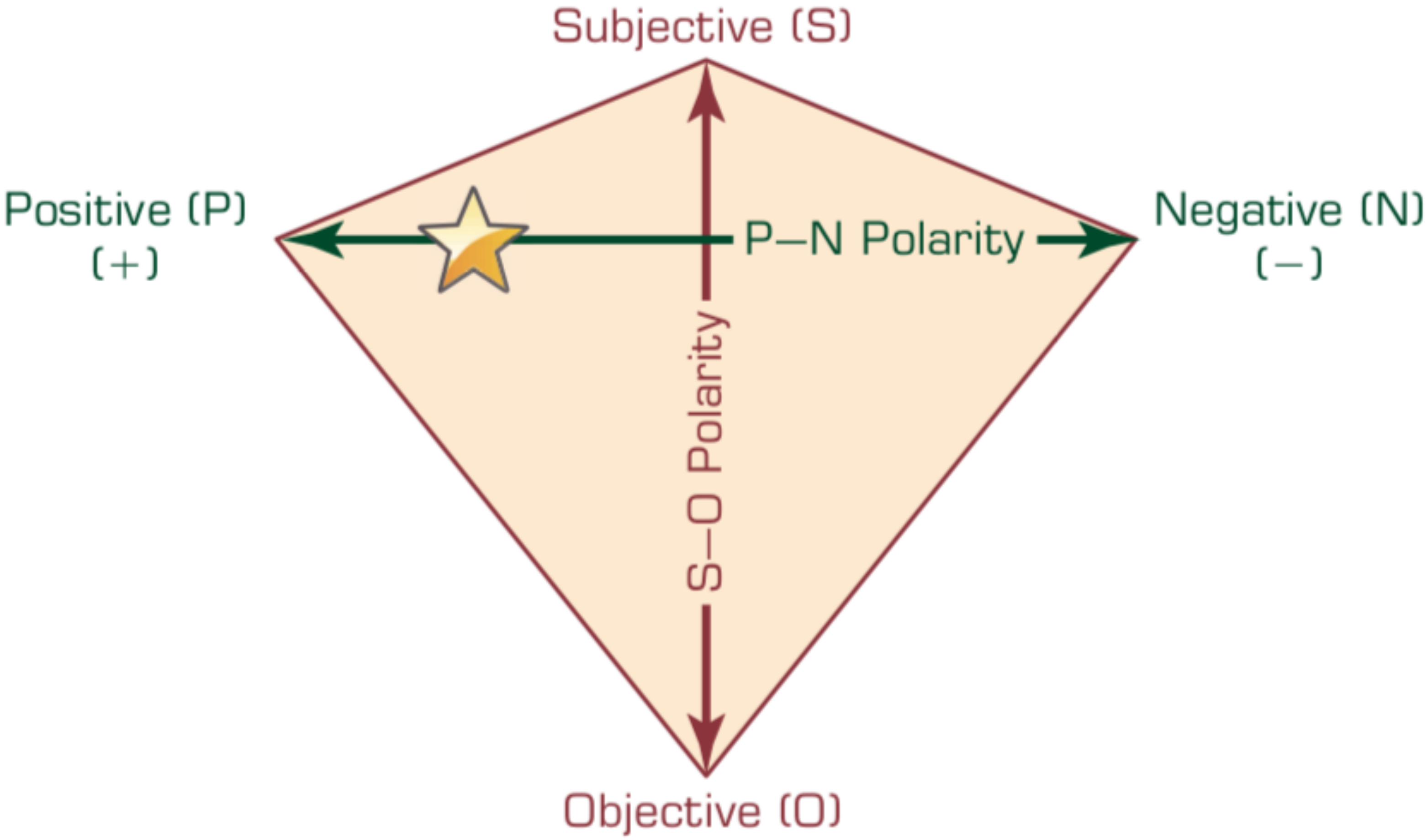
Source: Bing Liu (2011) , "Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data," Springer, 2nd Edition,

A Multistep Process to Sentiment Analysis


Textual Data
A statement
Step 1
Calculate the O–S Polarity
Lexicon
Is there a sentiment?
No
Yes
Yes
O–S Polarity measure
Step 2
Calculate the N–P Polarity of the sentiment
Lexicon
N–P Polarity
Record the Polarity, Strength, and the Target of the sentiment
Step 3
Identify the target for the sentiment
Target
Step 4
Tabulate & aggregate the sentiment analysis results

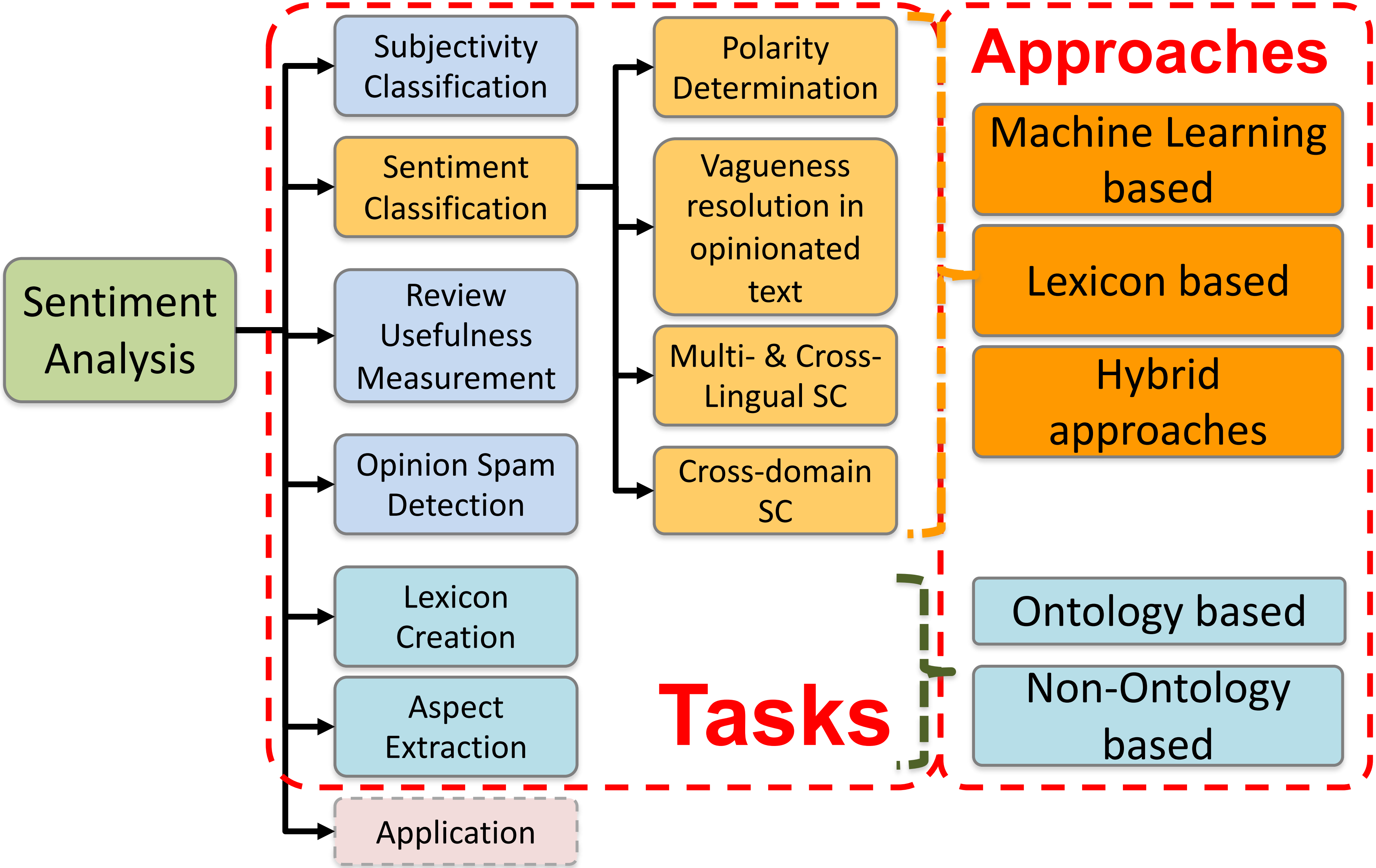
Source: Ramesh Sharda, Dursun Delen, and Efraim Turban (2017), Business Intelligence, Analytics, and Data Science: A Managerial Perspective, 4th Edition, Pearson

P–N Polarity and S–O Polarity Relationship



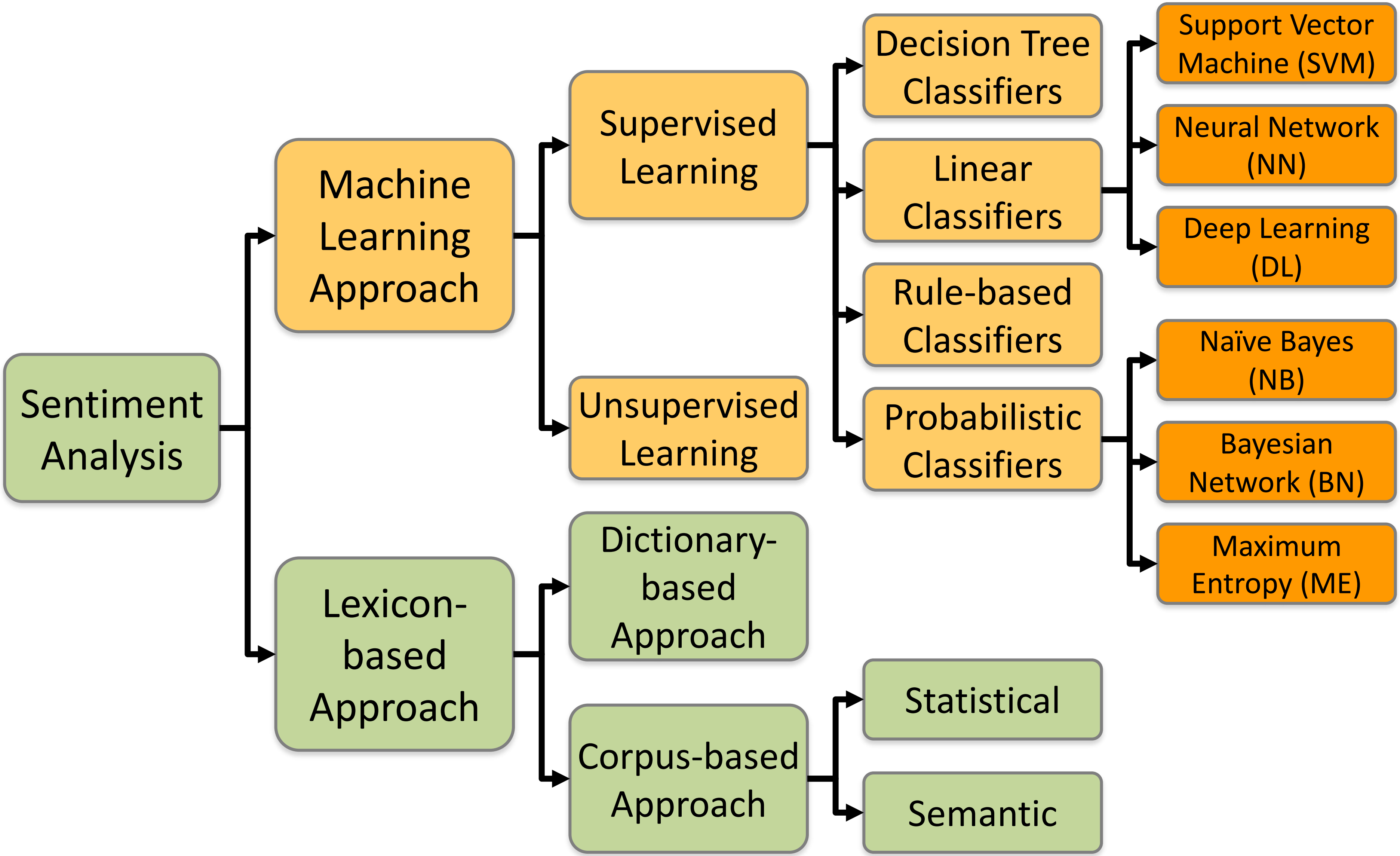
Source: Ramesh Sharda, Dursun Delen, and Efraim Turban (2017), Business Intelligence, Analytics, and Data Science: A Managerial Perspective, 4th Edition, Pearson

Sentiment Analysis



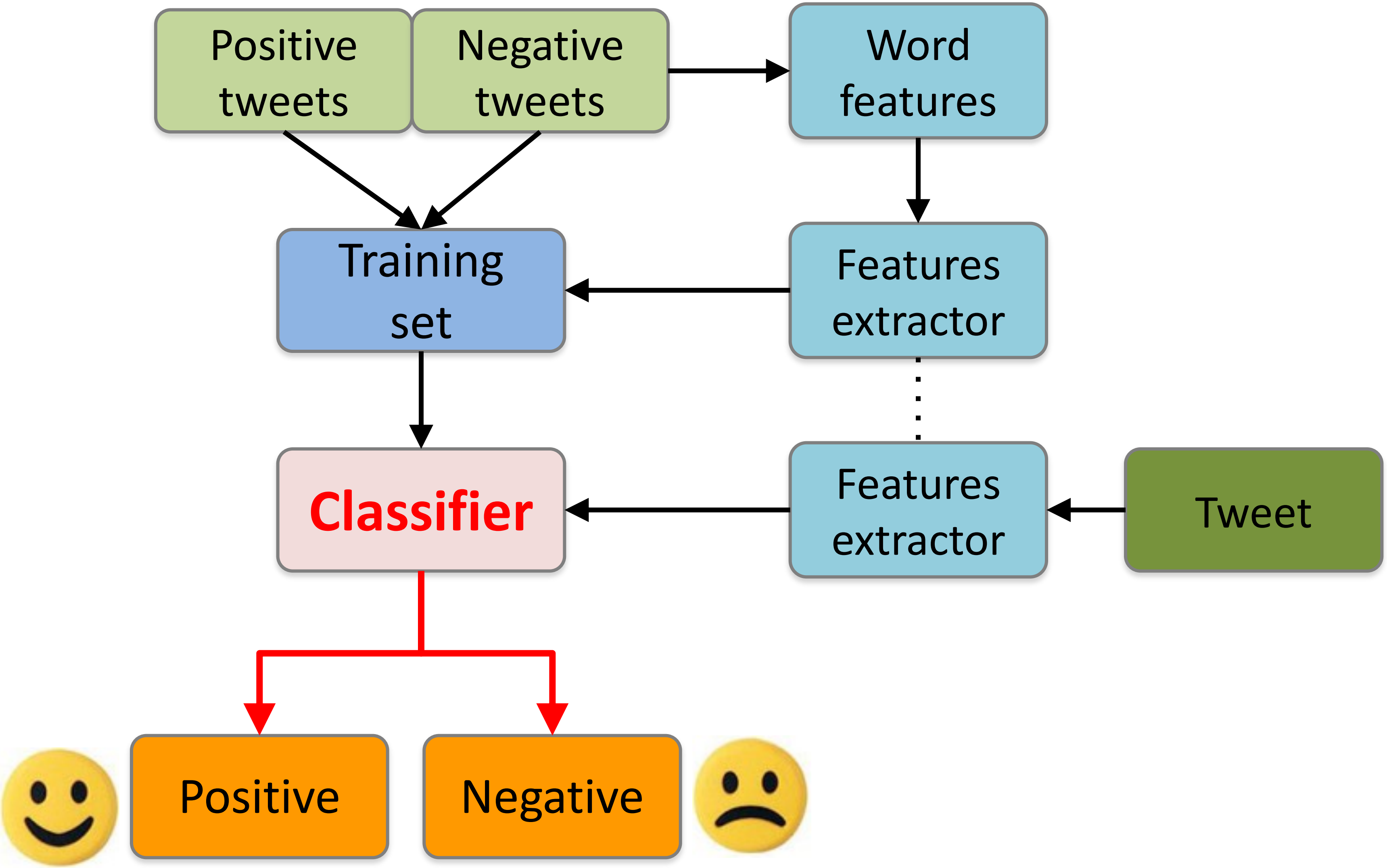
Source: Kumar Ravi and Vadlamani Ravi (2015), "A survey on opinion mining and sentiment analysis: tasks, approaches and applications." Knowledge-Based Systems, 89, pp.14-46.

Sentiment Classification Techniques



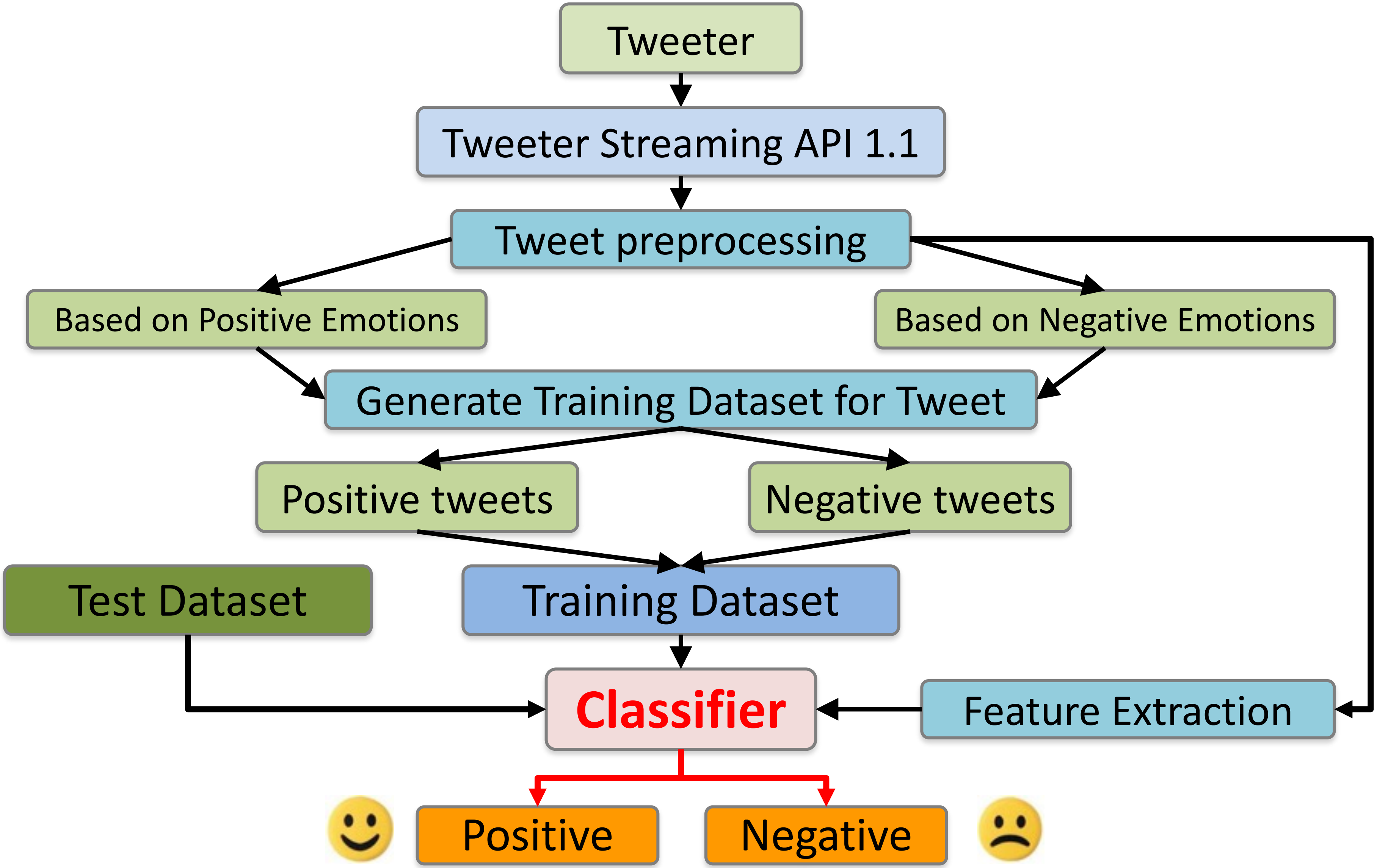
Source: Jesus Serrano-Guerrero, Jose A. Olivas, Francisco P. Romero, and Enrique Herrera-Viedma (2015), "Sentiment analysis: A review and comparative analysis of web services," Information Sciences, 311, pp. 18-38.

Sentiment Analysis Architecture



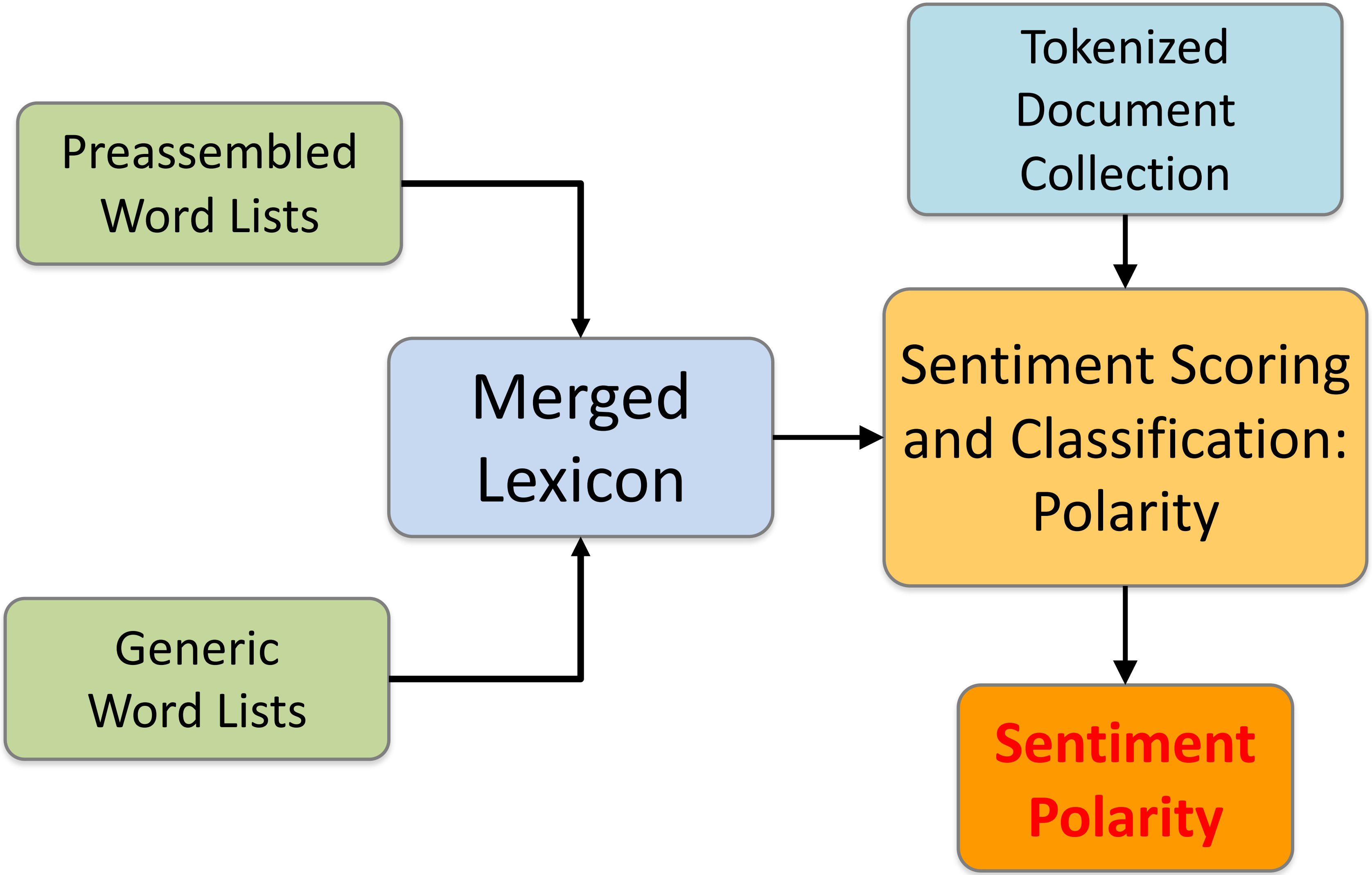
Vishal Kharde and Sheetal Sonawane (2016), "Sentiment Analysis of Twitter Data: A Survey of Techniques," International Journal of Computer Applications, Vol 139, No. 11, 2016. pp.5-15

Sentiment Classification Based on Emoticons



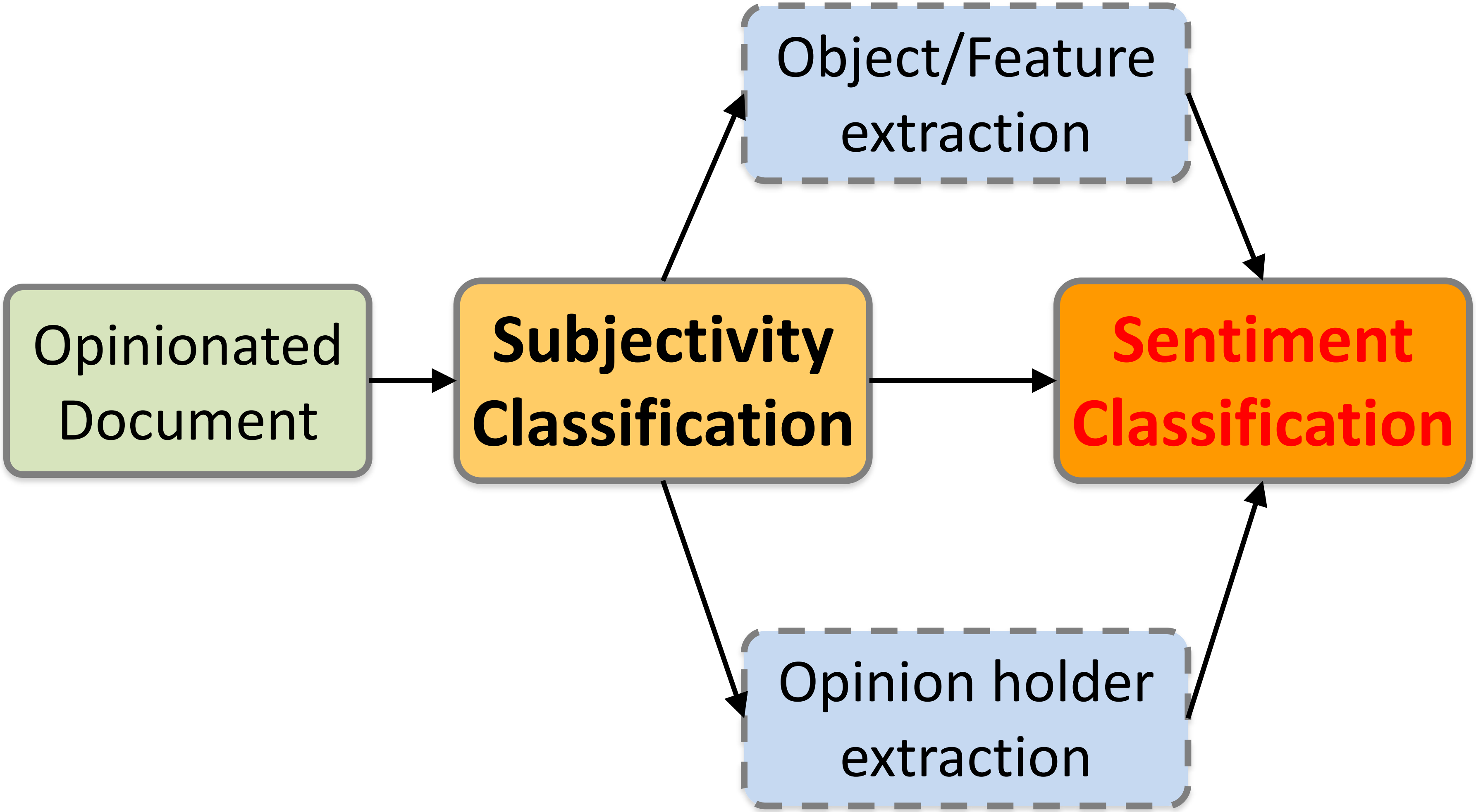
Vishal Kharde and Sheetal Sonawane (2016), "Sentiment Analysis of Twitter Data: A Survey of Techniques," International Journal of Computer Applications, Vol 139, No. 11, 2016. pp.5-15

Lexicon-Based Model


Preassembled Word Lists
Generic Word Lists
Merged Lexicon
Tokenized Document Collection
Sentiment Scoring and Classification: Polarity
Sentiment Polarity

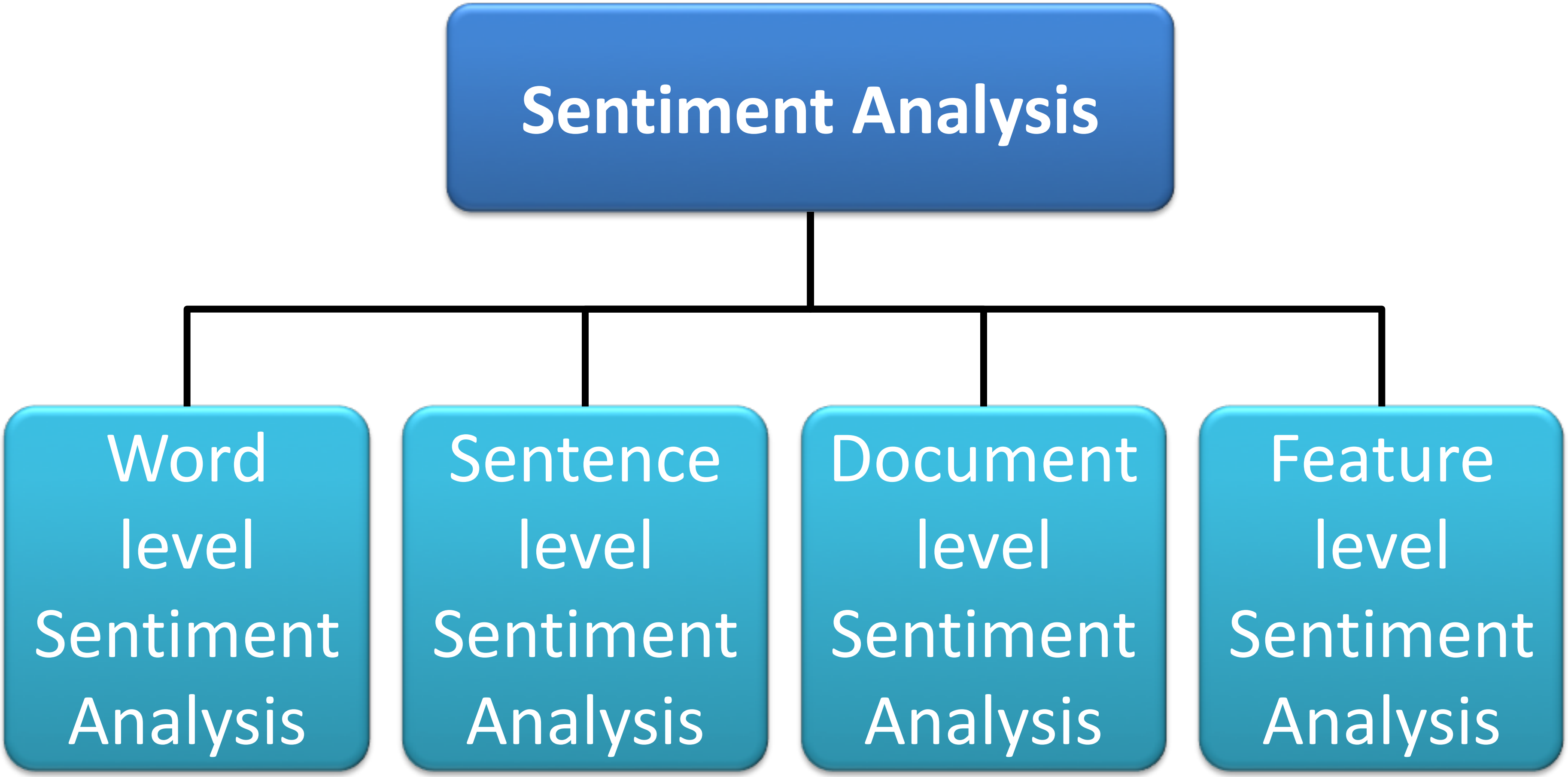
Vishal Kharde and Sheetal Sonawane (2016), "Sentiment Analysis of Twitter Data: A Survey of Techniques," International Journal of Computer Applications, Vol 139, No. 11, 2016. pp.5-15

Sentiment Analysis Tasks


Object/Feature extraction
Opinionated Document
Subjectivity Classification
Sentiment Classification
Opinion holder extraction

Vishal Kharde and Sheetal Sonawane (2016), "Sentiment Analysis of Twitter Data: A Survey of Techniques," International Journal of Computer Applications, Vol 139, No. 11, 2016. pp.5-15

Levels of Sentiment Analysis

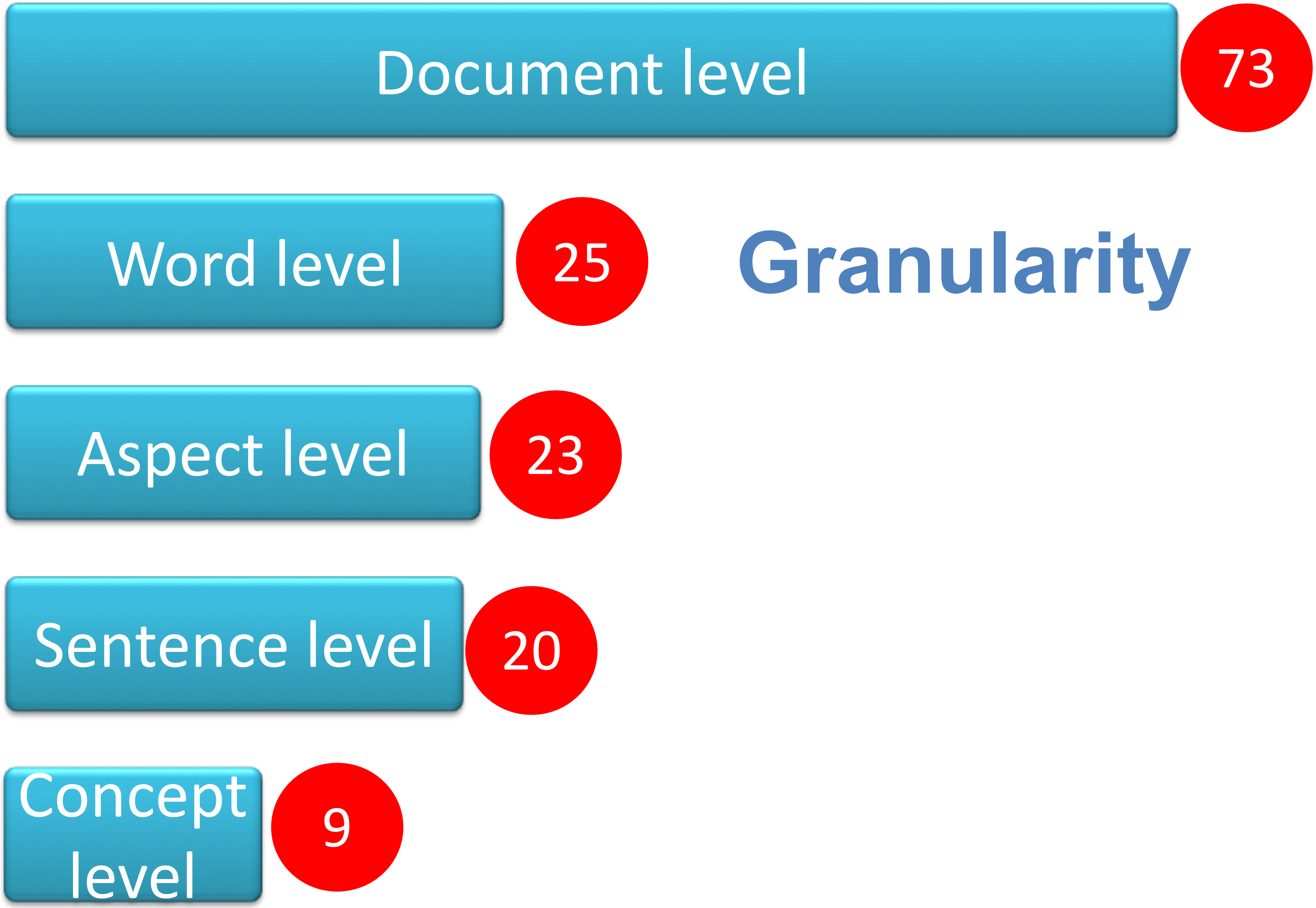


Sentiment Analysis

- Word level Sentiment Analysis
- Sentence level Sentiment Analysis
- Document level Sentiment Analysis
- Feature level Sentiment Analysis

Vishal Kharde and Sheetal Sonawane (2016), "Sentiment Analysis of Twitter Data: A Survey of Techniques," International Journal of Computer Applications, Vol 139, No. 11, 2016. pp.5-15

Levels of Sentiment Analysis



Document level 73

Word level 25

Aspect level 23

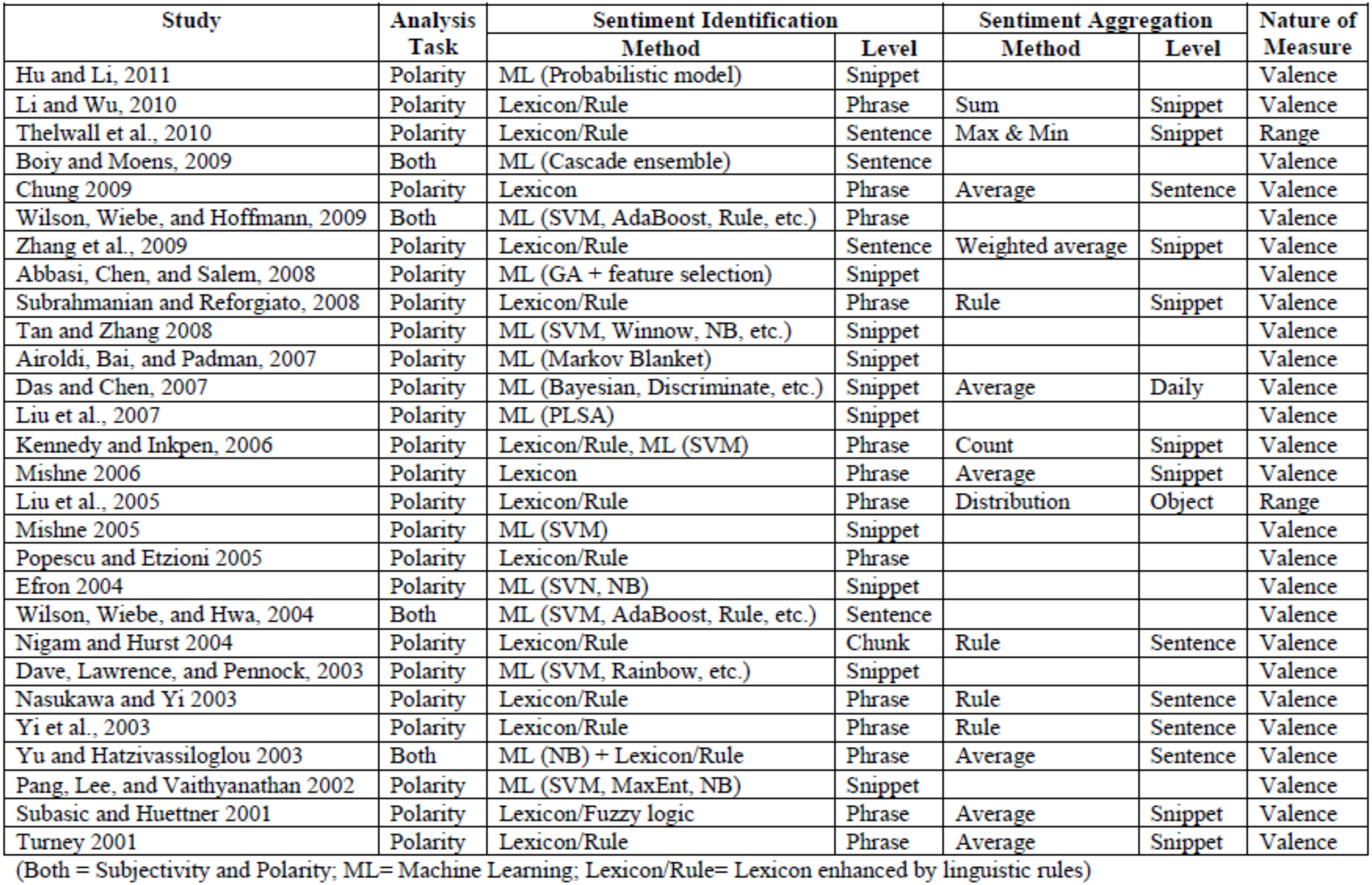
Sentence level 20

Concept level 9

Granularity

Source: Kumar Ravi and Vadlamani Ravi (2015), "A survey on opinion mining and sentiment analysis: tasks, approaches and applications." Knowledge-Based Systems, 89, pp.14-46.

A Brief Summary of Sentiment Analysis Methods



Study	Analysis Task	Sentiment Identification		Sentiment Aggregation		Nature of Measure
		Method	Level	Method	Level	
Hu and Li, 2011	Polarity	ML (Probabilistic model)	Snippet			Valence
Li and Wu, 2010	Polarity	Lexicon/Rule	Phrase	Sum	Snippet	Valence
Thelwall et al., 2010	Polarity	Lexicon/Rule	Sentence	Max & Min	Snippet	Range
Boiy and Moens, 2009	Both	ML (Cascade ensemble)	Sentence			Valence
Chung 2009	Polarity	Lexicon	Phrase	Average	Sentence	Valence
Wilson, Wiebe, and Hoffmann, 2009	Both	ML (SVM, AdaBoost, Rule, etc.)	Phrase			Valence
Zhang et al., 2009	Polarity	Lexicon/Rule	Sentence	Weighted average	Snippet	Valence
Abbasi, Chen, and Salem, 2008	Polarity	ML (GA + feature selection)	Snippet			Valence
Subrahmanian and Reforgiato, 2008	Polarity	Lexicon/Rule	Phrase	Rule	Snippet	Valence
Tan and Zhang 2008	Polarity	ML (SVM, Winnow, NB, etc.)	Snippet			Valence
Airoldi, Bai, and Padman, 2007	Polarity	ML (Markov Blanket)	Snippet			Valence
Das and Chen, 2007	Polarity	ML (Bayesian, Discriminate, etc.)	Snippet	Average	Daily	Valence
Liu et al., 2007	Polarity	ML (PLSA)	Snippet			Valence
Kennedy and Inkpen, 2006	Polarity	Lexicon/Rule, ML (SVM)	Phrase	Count	Snippet	Valence
Mishne 2006	Polarity	Lexicon	Phrase	Average	Snippet	Valence
Liu et al., 2005	Polarity	Lexicon/Rule	Phrase	Distribution	Object	Range
Mishne 2005	Polarity	ML (SVM)	Snippet			Valence
Popescu and Etzioni 2005	Polarity	Lexicon/Rule	Phrase			Valence
Efron 2004	Polarity	ML (SVN, NB)	Snippet			Valence
Wilson, Wiebe, and Hwa, 2004	Both	ML (SVM, AdaBoost, Rule, etc.)	Sentence			Valence
Nigam and Hurst 2004	Polarity	Lexicon/Rule	Chunk	Rule	Sentence	Valence
Dave, Lawrence, and Pennock, 2003	Polarity	ML (SVM, Rainbow, etc.)	Snippet			Valence
Nasukawa and Yi 2003	Polarity	Lexicon/Rule	Phrase	Rule	Sentence	Valence
Yi et al., 2003	Polarity	Lexicon/Rule	Phrase	Rule	Sentence	Valence
Yu and Hatzivassiloglou 2003	Both	ML (NB) + Lexicon/Rule	Phrase	Average	Sentence	Valence
Pang, Lee, and Vaithyanathan 2002	Polarity	ML (SVM, MaxEnt, NB)	Snippet			Valence
Subasic and Huettner 2001	Polarity	Lexicon/Fuzzy logic	Phrase	Average	Snippet	Valence
Turney 2001	Polarity	Lexicon/Rule	Phrase	Average	Snippet	Valence

(Both = Subjectivity and Polarity; ML= Machine Learning; Lexicon/Rule= Lexicon enhanced by linguistic rules)

Source: Zhang, Z., Li, X., and Chen, Y. (2012), "Deciphering word-of-mouth in social media: Text-based metrics of consumer reviews," ACM Trans. Manage. Inf. Syst. (3:1) 2012, pp 1-23.,

Word-of-Mouth (WOM)

- “This book is the best written documentary thus far, yet sadly, there is no soft cover edition.”

- “This book is the best written documentary thus far, yet sadly, there is no soft cover edition.”

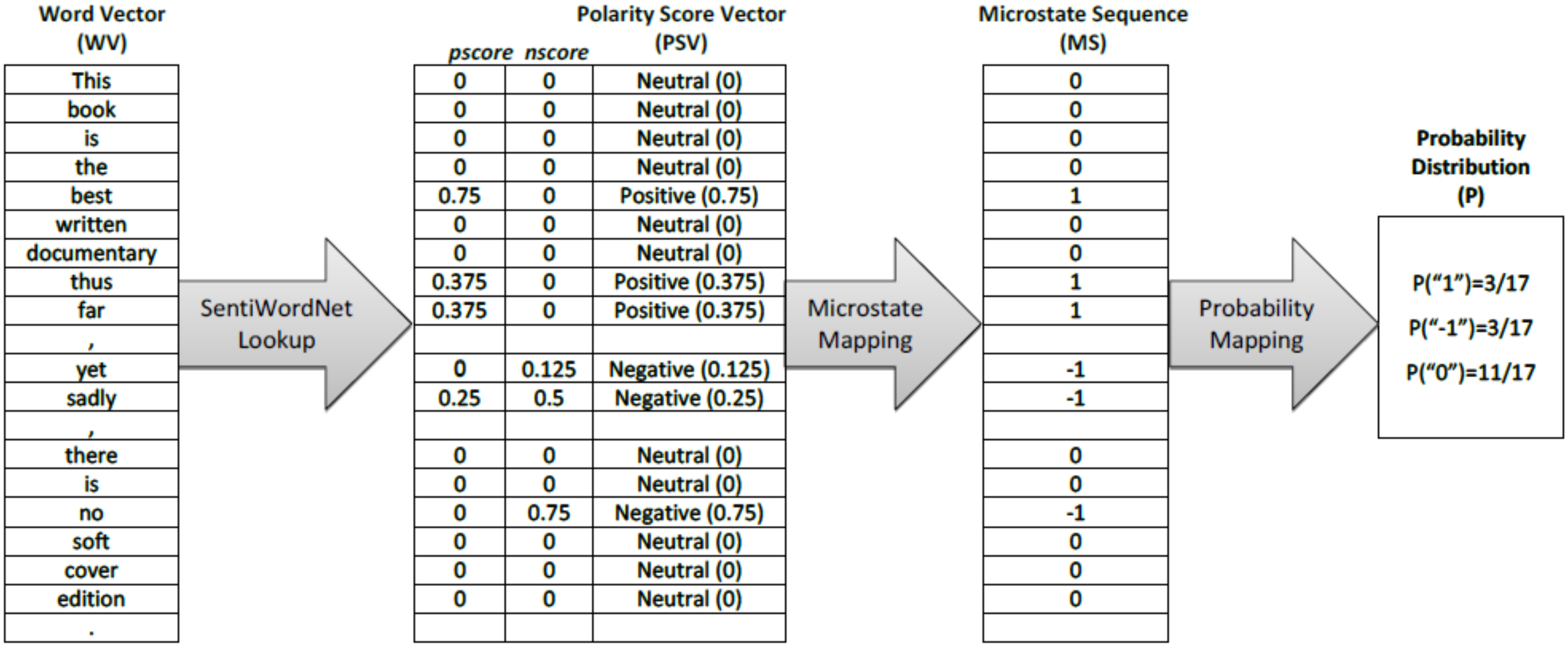
Source: Zhang, Z., Li, X., and Chen, Y. (2012), "Deciphering word-of-mouth in social media: Text-based metrics of consumer reviews," ACM Trans. Manage. Inf. Syst. (3:1) 2012, pp 1-23.,

This
book
is
the
best
written
documentary
thus
far
,
yet
sadly
,
there
is
no
soft
cover
edition
.

Word	POS
This	DT
book	NN
is	VBZ
the	DT
best	JJS
written	VBN
documentary	NN
thus	RB
far	RB
,	,
yet	RB
sadly	RB
,	,
there	EX
is	VBZ
no	DT
soft	JJ
cover	NN
edition	NN
.	.

Source: Zhang, Z., Li, X., and Chen, Y. (2012), "Deciphering word-of-mouth in social media: Text-based metrics of consumer reviews," ACM Trans. Manage. Inf. Syst. (3:1) 2012, pp 1-23.,

Conversion of text representation



Word Vector (WV)
This
book
is
the
best
written
documentary
thus
far
,
yet
sadly
,
there
is
no
soft
cover
edition
.

pscore	*nscore*	Polarity Score Vector (PSV)
0	0	Neutral (0)
0	0	Neutral (0)
0	0	Neutral (0)
0	0	Neutral (0)
0.75	0	Positive (0.75)
0	0	Neutral (0)
0	0	Neutral (0)
0.375	0	Positive (0.375)
0.375	0	Positive (0.375)
0	0.125	Negative (0.125)
0.25	0.5	Negative (0.25)
0	0	Neutral (0)
0	0	Neutral (0)
0	0.75	Negative (0.75)
0	0	Neutral (0)
0	0	Neutral (0)
0	0	Neutral (0)

Microstate Sequence (MS)
0
0
0
0
1
0
0
1
1
-1
-1
0
0
-1
0
0
0

Probability Distribution (P)

P("1")=3/17

P("-1")=3/17

P("0")=11/17

Source: Zhang, Z., Li, X., and Chen, Y. (2012), "Deciphering word-of-mouth in social media: Text-based metrics of consumer reviews," ACM Trans. Manage. Inf. Syst. (3:1) 2012, pp 1-23.,

Example of SentiWordNet

POS	ID	PosScore	NegScore	SynsetTerms	Gloss
a	00217728	0.75	0	beautiful#1	delighting the senses or exciting intellectual or emotional admiration; "a beautiful child"; "beautiful country"; "a beautiful painting"; "a beautiful theory"; "a beautiful party“
a	00227507	0.75	0	best#1	(superlative of `good') having the most positive qualities; "the best film of the year"; "the best solution"; "the best time for planting"; "wore his best suit“
r	00042614	0	0.625	unhappily#2 sadly#1	in an unfortunate way; "sadly he died before he could see his grandchild“
r	00093270	0	0.875	woefully#1 sadly#3 lamentably#1 deplorably#1	in an unfortunate or deplorable manner; "he was sadly neglected"; "it was woefully inadequate“
r	00404501	0	0.25	sadly#2	with sadness; in a sad manner; "`She died last night,' he said sadly"

SenticNet

The car is very old but it is rather not expensive.

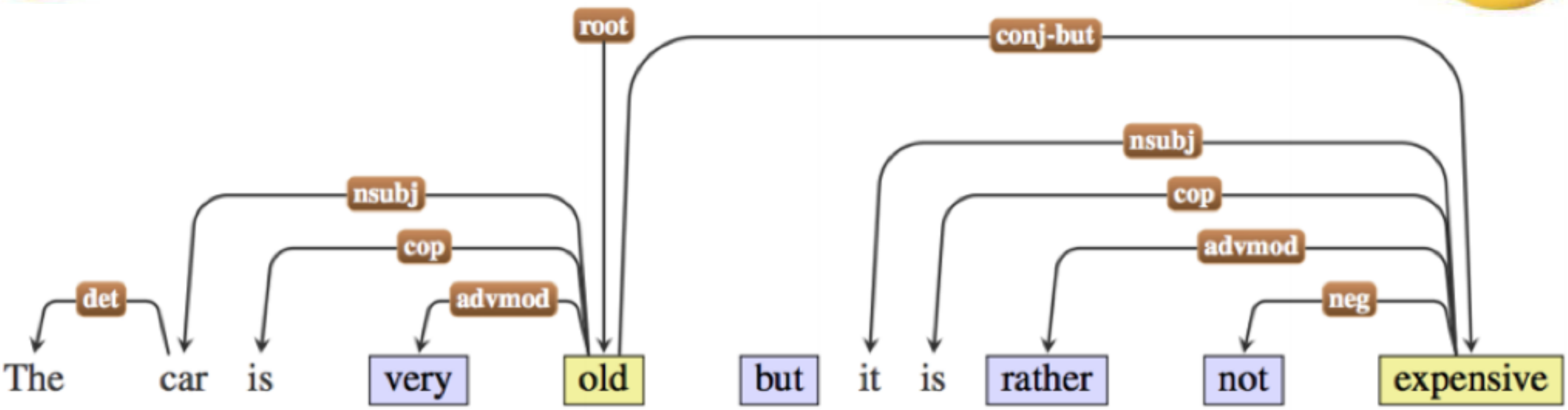
The car is very old but it is rather not expensive.

The car is very old but it is rather not expensive.

Source: Cambria, Erik, Soujanya Poria, Rajiv Bajpai, and Björn Schuller. "SenticNet 4: A semantic resource for sentiment analysis based on conceptual primitives." In *the 26th International Conference on Computational Linguistics (COLING), Osaka*. 2016.

Polarity Detection with SenticNet

🙂 🙁

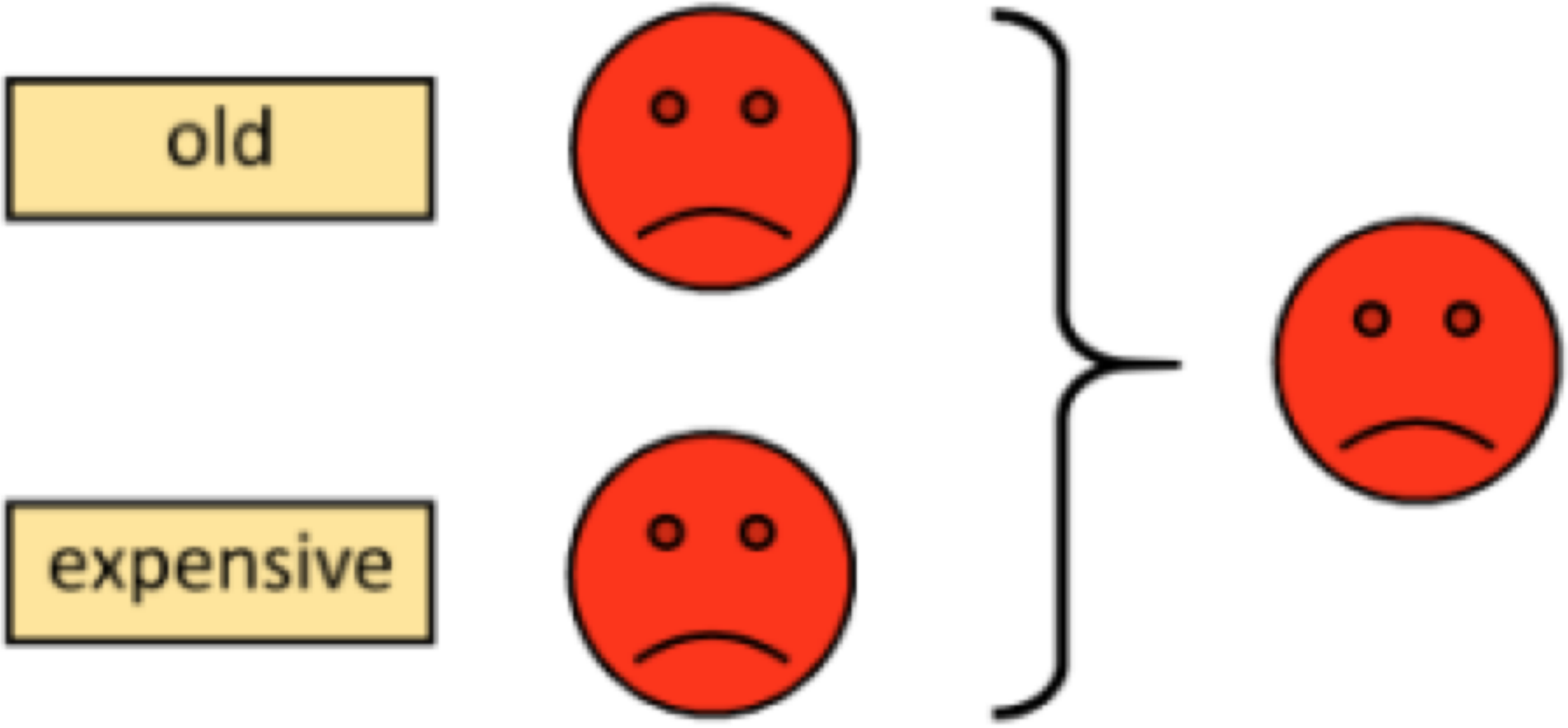


The car is very old but it is rather not expensive.

The car is very old but it is rather not expensive.

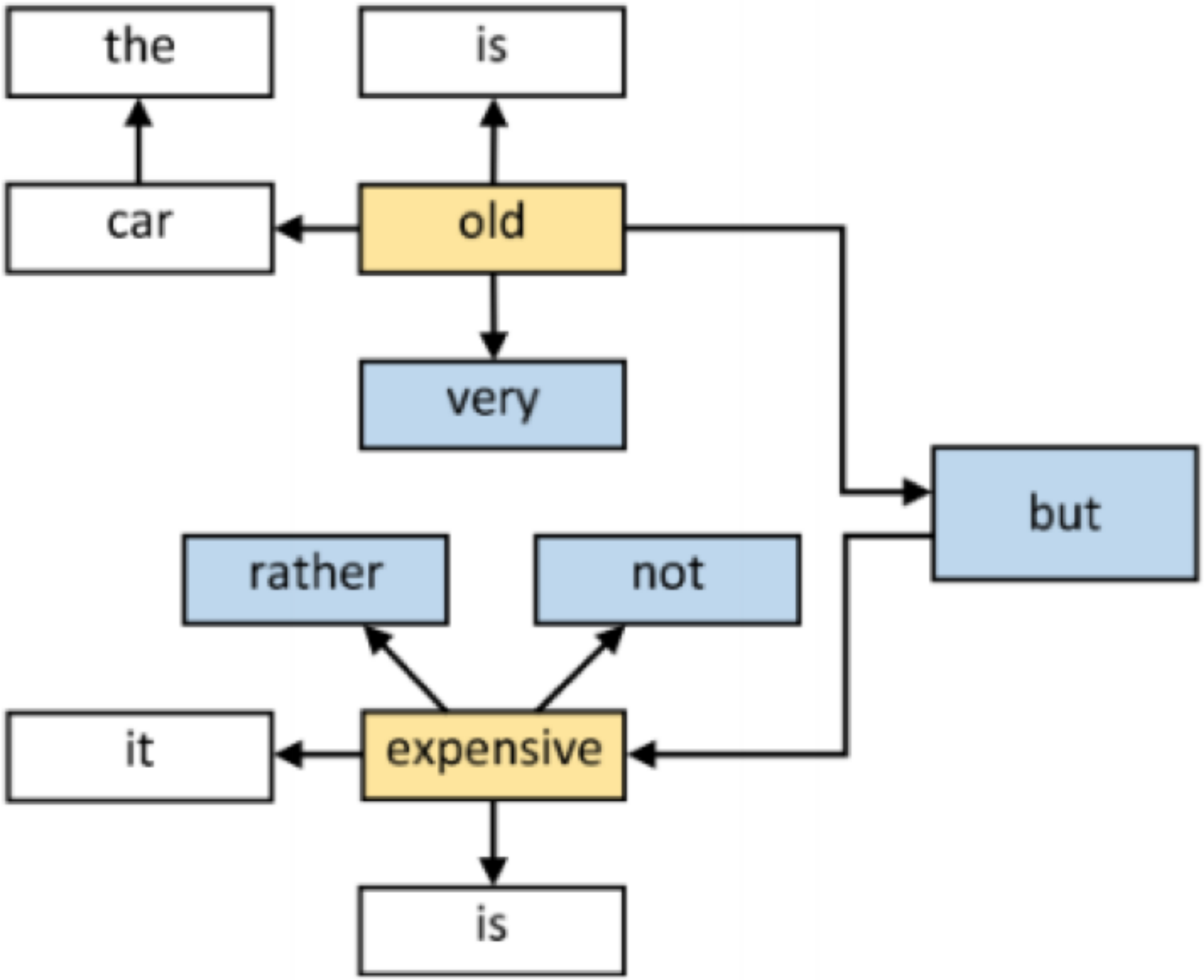
Source: Cambria, Erik, Soujanya Poria, Rajiv Bajpai, and Björn Schuller. "SenticNet 4: A semantic resource for sentiment analysis based on conceptual primitives." In *the 26th International Conference on Computational Linguistics (COLING), Osaka*. 2016.

Polarity Detection with SenticNet



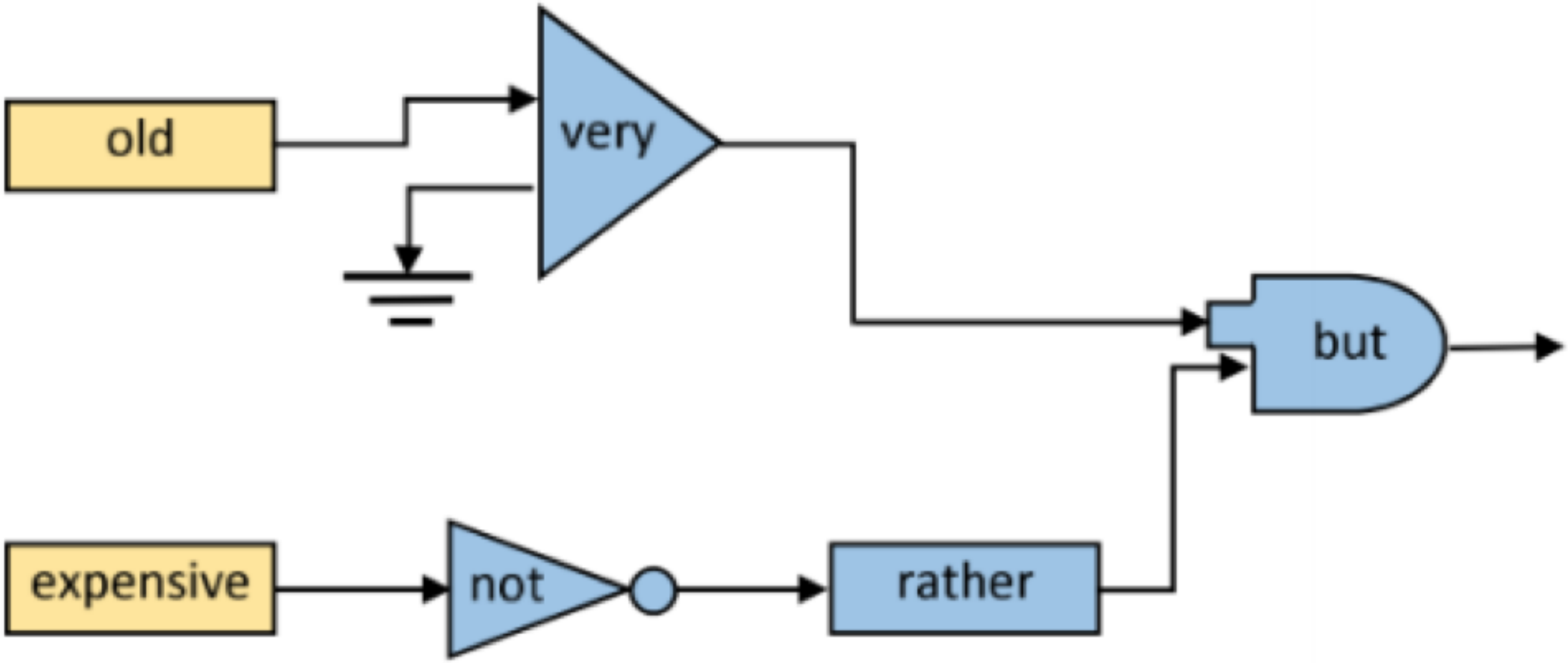
Source: Cambria, Erik, Soujanya Poria, Rajiv Bajpai, and Björn Schuller. "SenticNet 4: A semantic resource for sentiment analysis based on conceptual primitives." In *the 26th International Conference on Computational Linguistics (COLING), Osaka*. 2016.

Polarity Detection with SenticNet



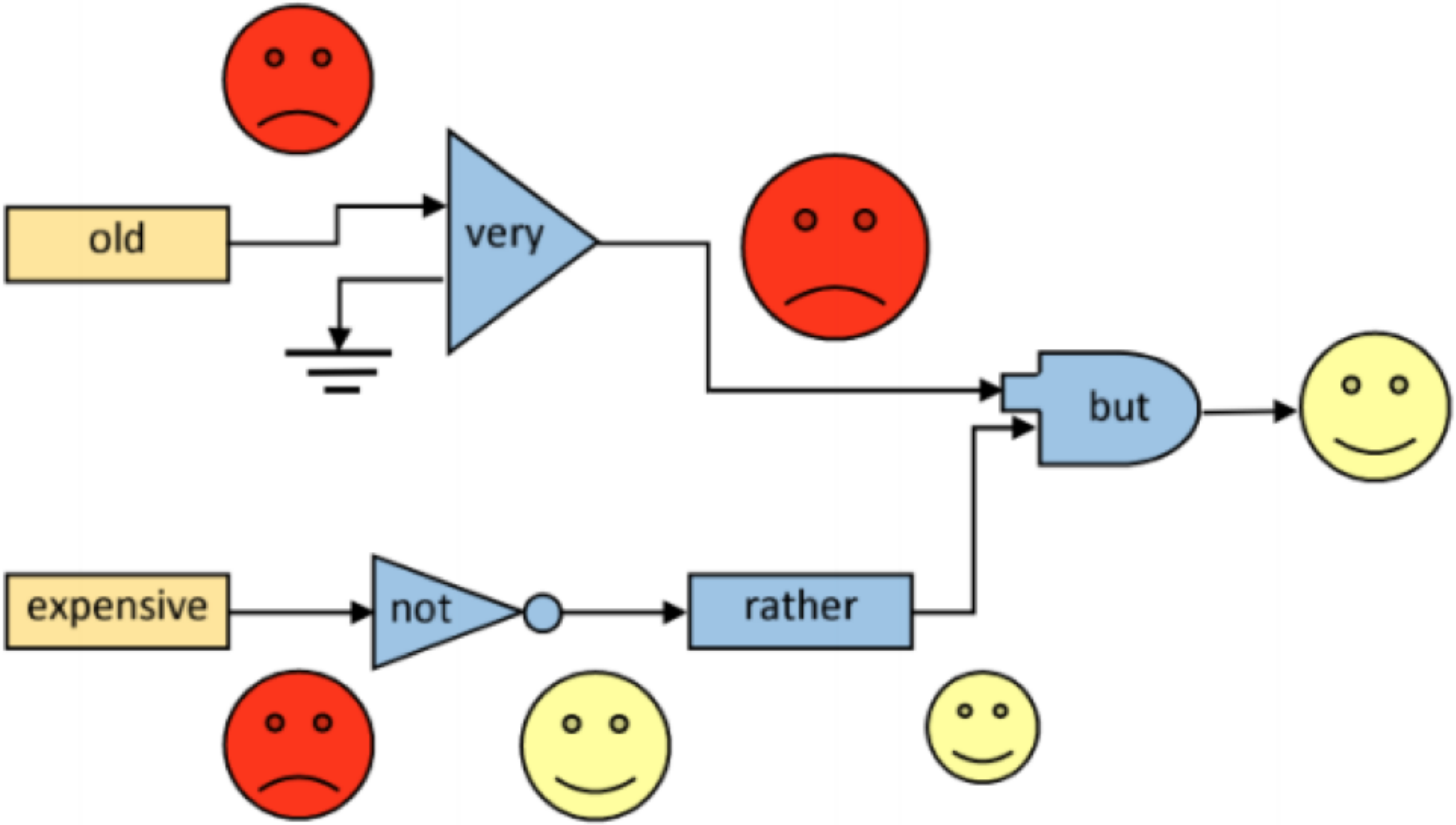
Source: Cambria, Erik, Soujanya Poria, Rajiv Bajpai, and Björn Schuller. "SenticNet 4: A semantic resource for sentiment analysis based on conceptual primitives." In *the 26th International Conference on Computational Linguistics (COLING), Osaka*. 2016.

Polarity Detection with SenticNet



Source: Cambria, Erik, Soujanya Poria, Rajiv Bajpai, and Björn Schuller. "SenticNet 4: A semantic resource for sentiment analysis based on conceptual primitives." In *the 26th International Conference on Computational Linguistics (COLING), Osaka*. 2016.

Polarity Detection with SenticNet



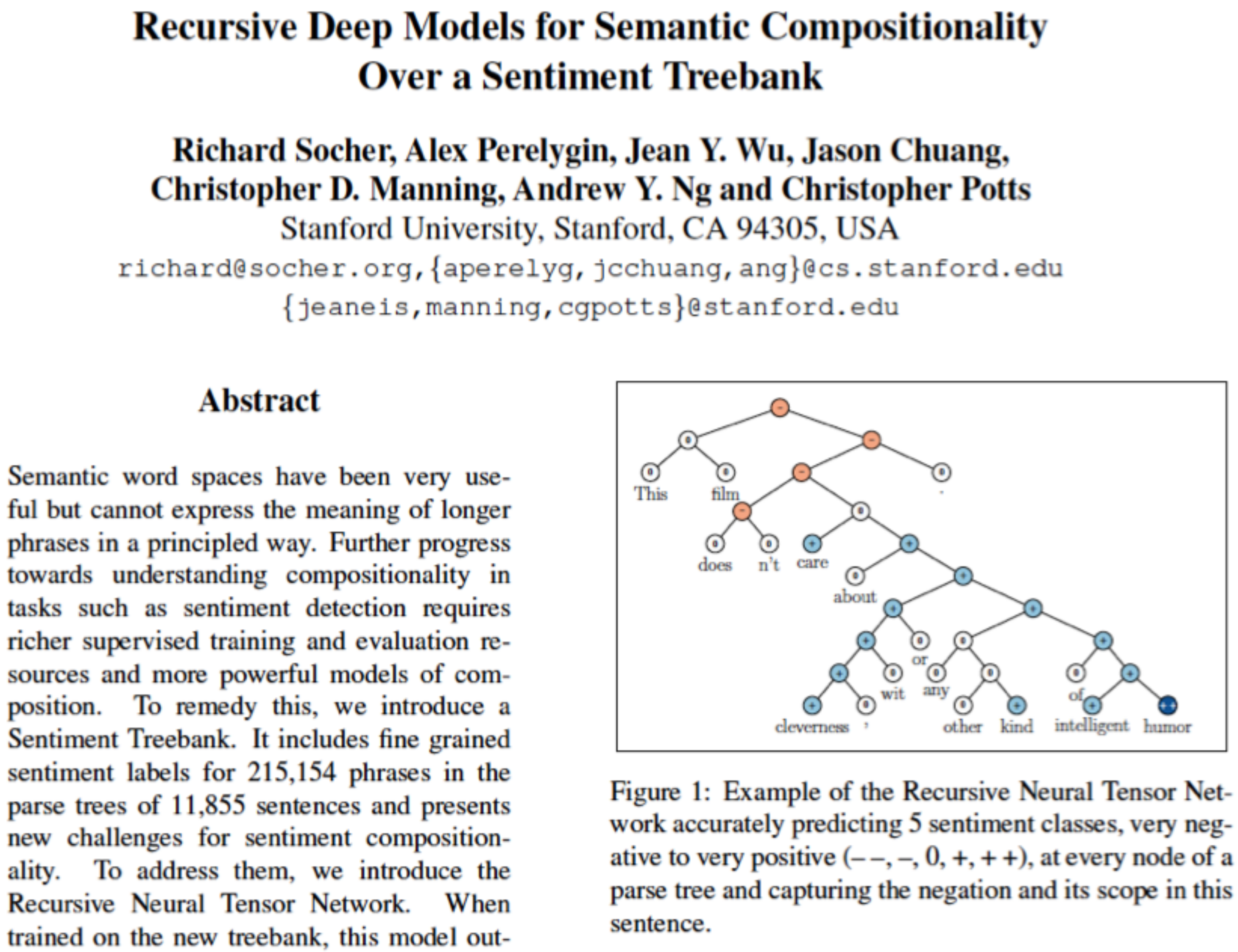
Source: Cambria, Erik, Soujanya Poria, Rajiv Bajpai, and Björn Schuller. "SenticNet 4: A semantic resource for sentiment analysis based on conceptual primitives." In *the 26th International Conference on Computational Linguistics (COLING), Osaka*. 2016.

Evaluation of Text Mining and Sentiment Analysis

- Evaluation of Information Retrieval
- Evaluation of Classification Model (Prediction)
  - Accuracy
  - Precision
  - Recall
  - F-score

Deep Learning for Sentiment Analytics

Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank



Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank

Richard Socher, Alex Perelygin, Jean Y. Wu, Jason Chuang, Christopher D. Manning, Andrew Y. Ng and Christopher Potts
Stanford University, Stanford, CA 94305, USA
richard@socher.org,{aperelyg,jcchuang,ang}@cs.stanford.edu
{jeaneis,manning,cgpotts}@stanford.edu

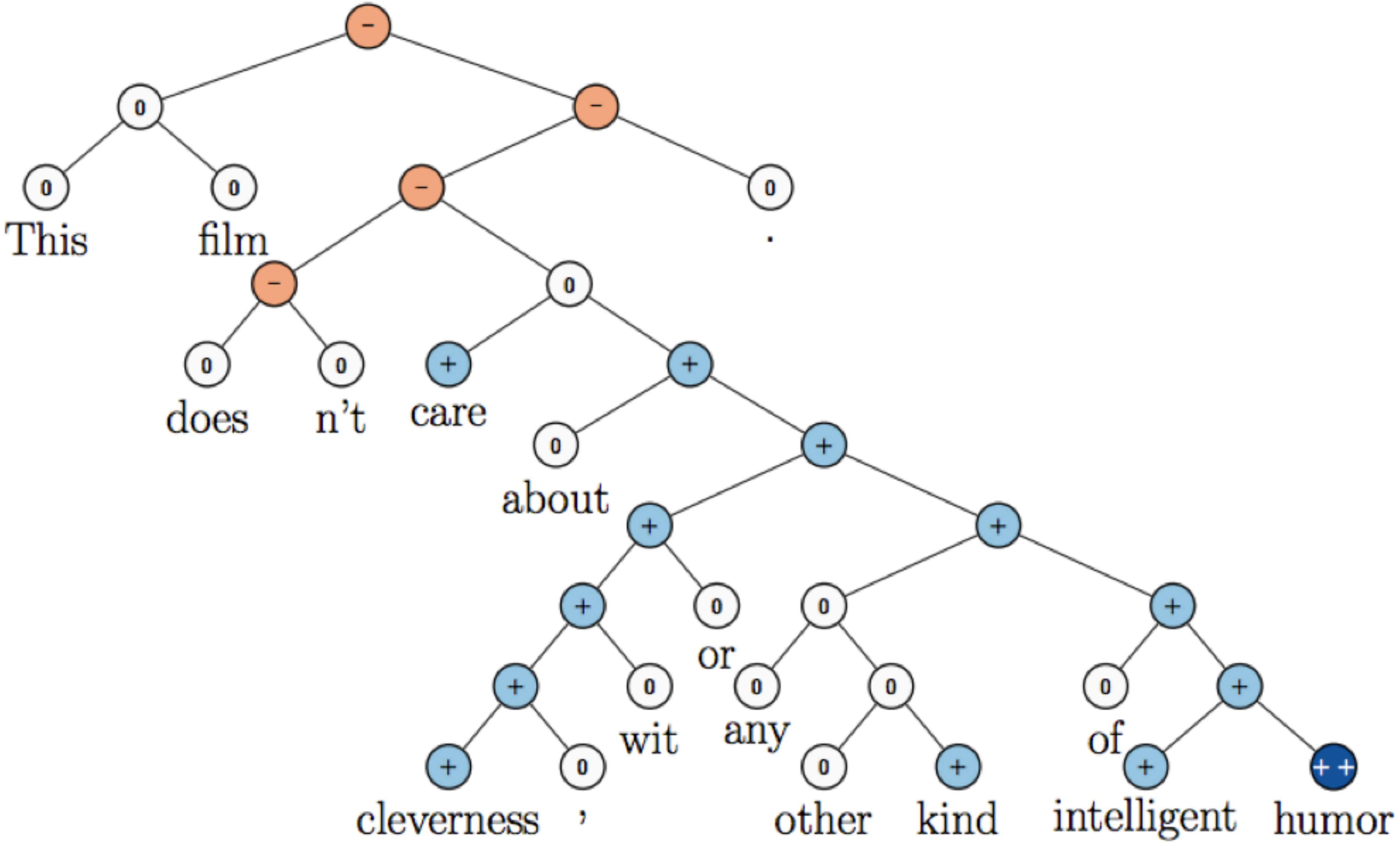
Abstract

Semantic word spaces have been very useful but cannot express the meaning of longer phrases in a principled way. Further progress towards understanding compositionality in tasks such as sentiment detection requires richer supervised training and evaluation resources and more powerful models of composition. To remedy this, we introduce a Sentiment Treebank. It includes fine grained sentiment labels for 215,154 phrases in the parse trees of 11,855 sentences and presents new challenges for sentiment compositionality. To address them, we introduce the Recursive Neural Tensor Network. When trained on the new treebank, this model out-

Figure 1: Example of the Recursive Neural Tensor Network accurately predicting 5 sentiment classes, very negative to very positive (– –, –, 0, +, + +), at every node of a parse tree and capturing the negation and its scope in this sentence.

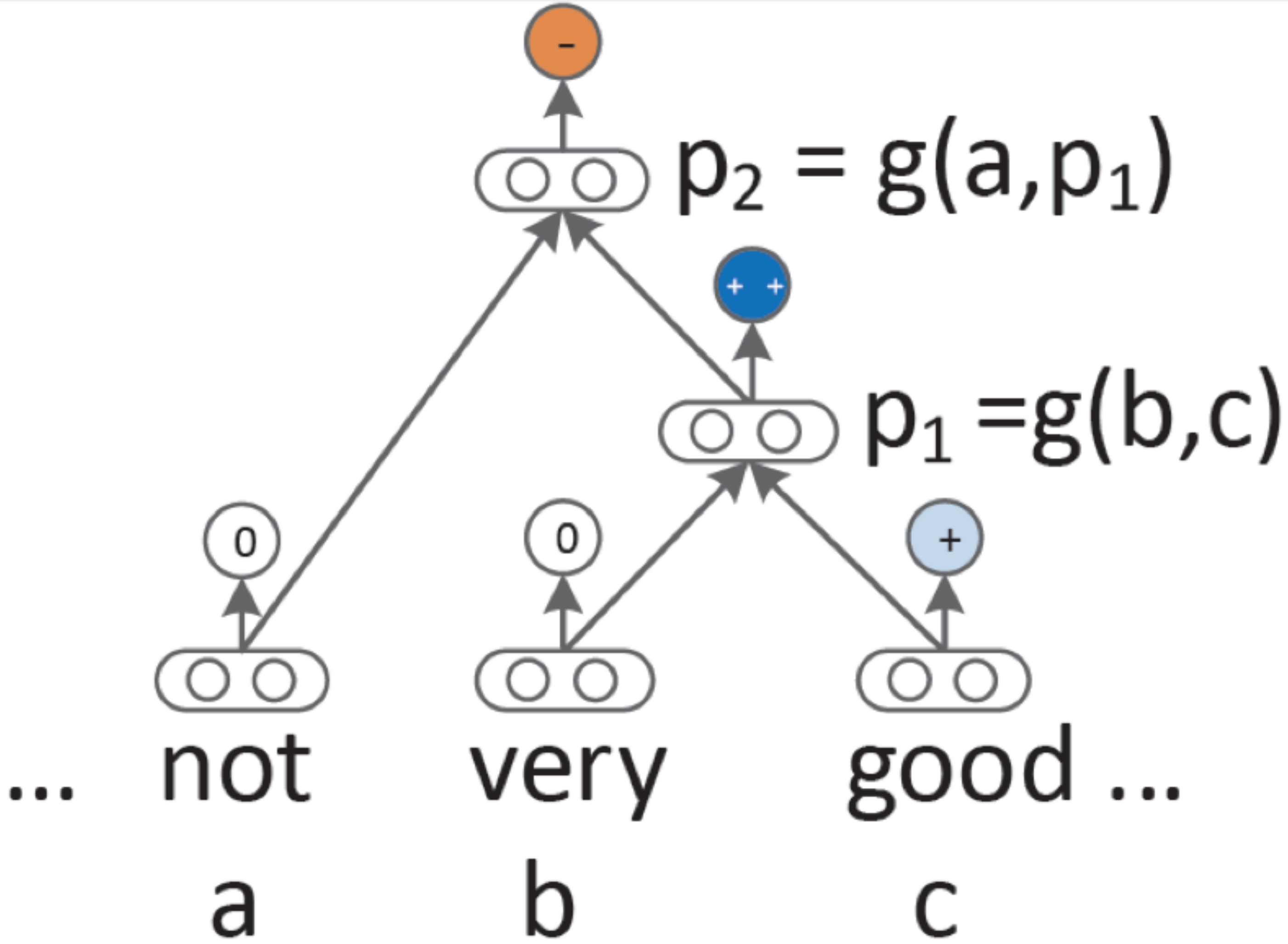
Source: Richard Socher et al. (2013) "Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank", EMNLP 2013

Recursive Neural Tensor Network (RNTN)



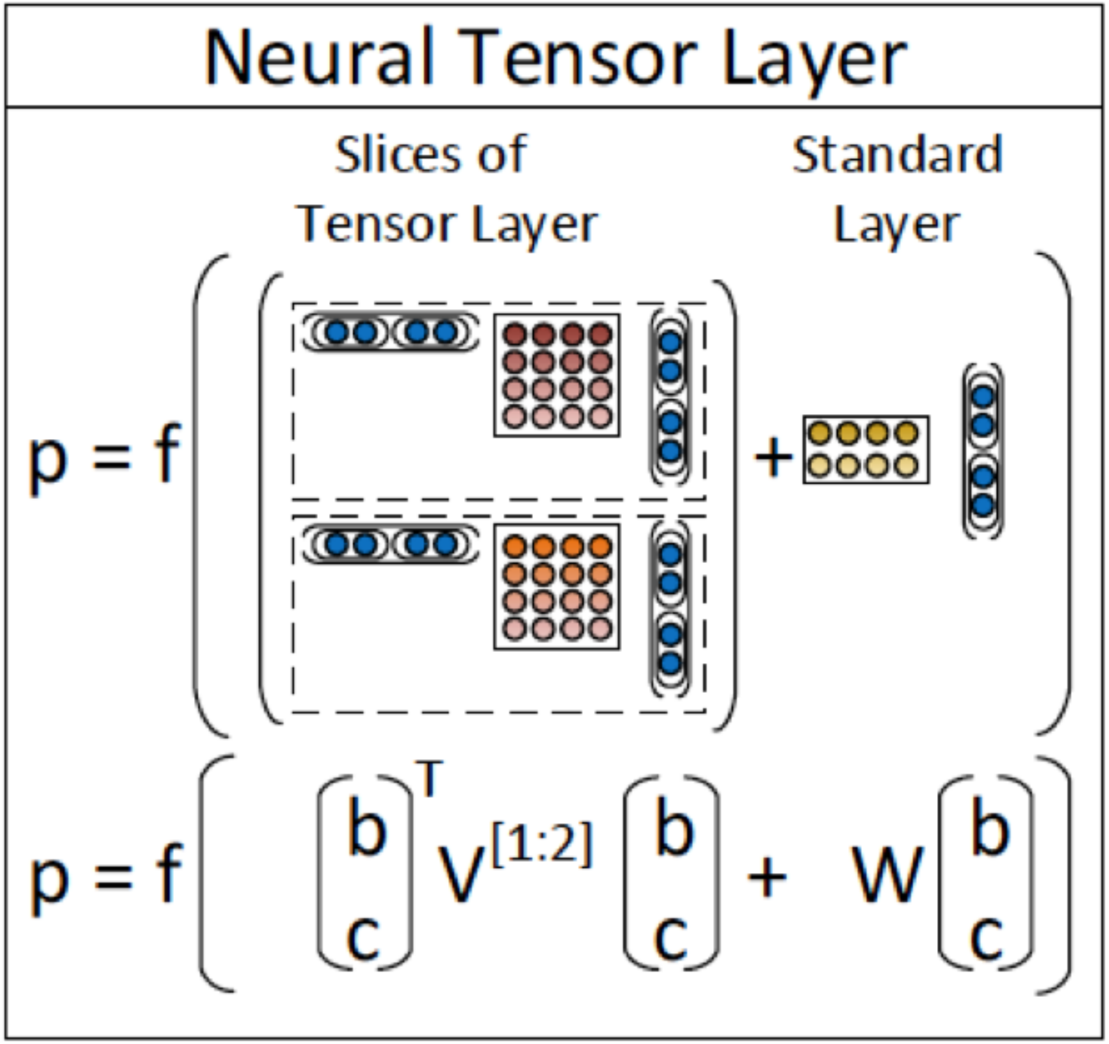
Source: Richard Socher et al. (2013) "Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank", EMNLP 2013

Recursive Neural Network (RNN) models for sentiment



Source: Richard Socher et al. (2013) "Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank", EMNLP 2013

Recursive Neural Tensor Network (RNTN)


Neural Tensor Layer
Slices of Tensor Layer
Standard Layer
p = f
+
$p = f\left(\begin{bmatrix} b \\ c \end{bmatrix}^T V^{[1:2]} \begin{bmatrix} b \\ c \end{bmatrix} + W \begin{bmatrix} b \\ c \end{bmatrix} \right)$

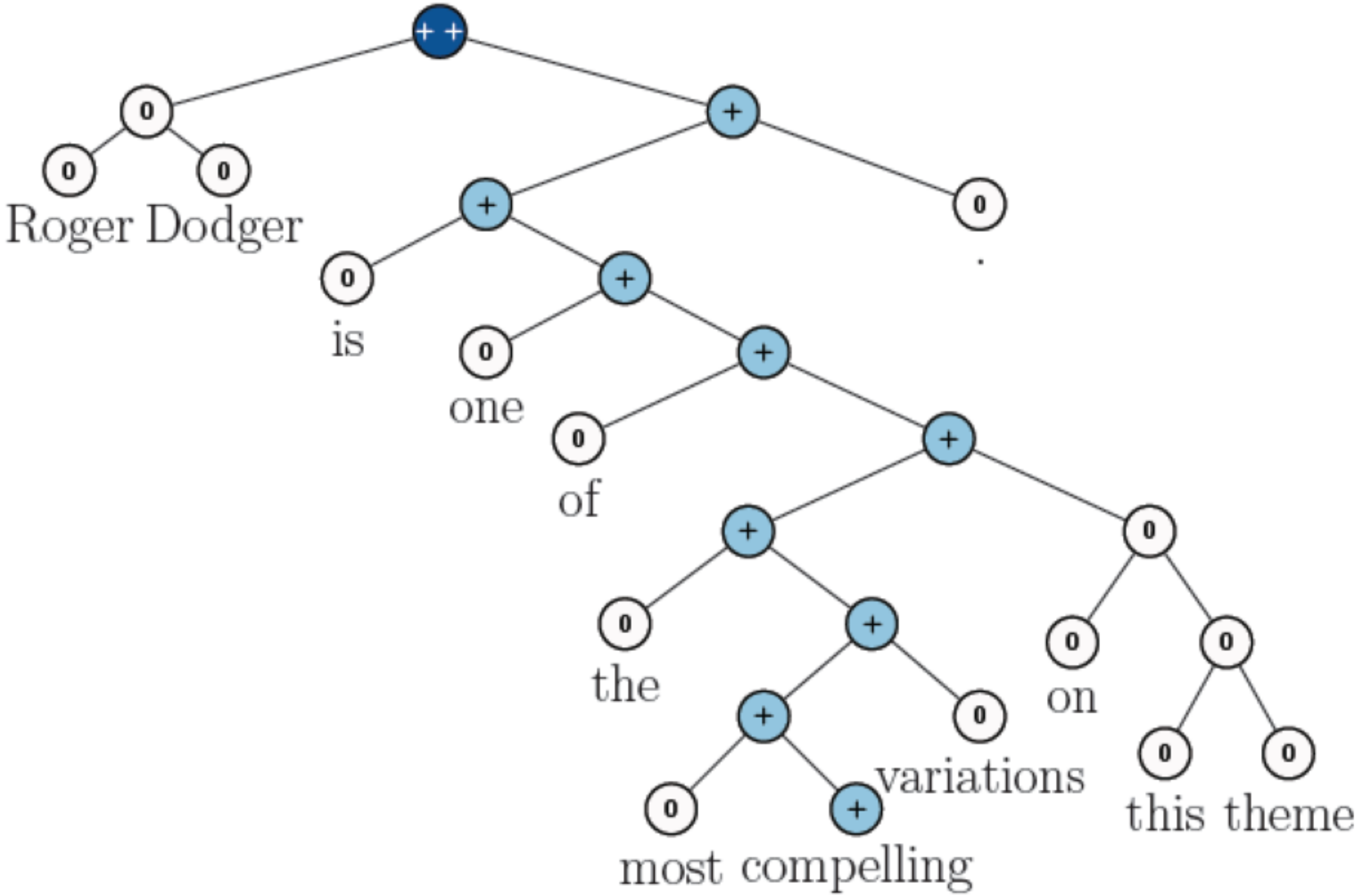
Source: Richard Socher et al. (2013) "Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank", EMNLP 2013

Roger Dodger is one of the most compelling variations on this theme.

Roger Dodger is one of the least compelling variations on this theme.

Source: Richard Socher et al. (2013) "Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank", EMNLP 2013

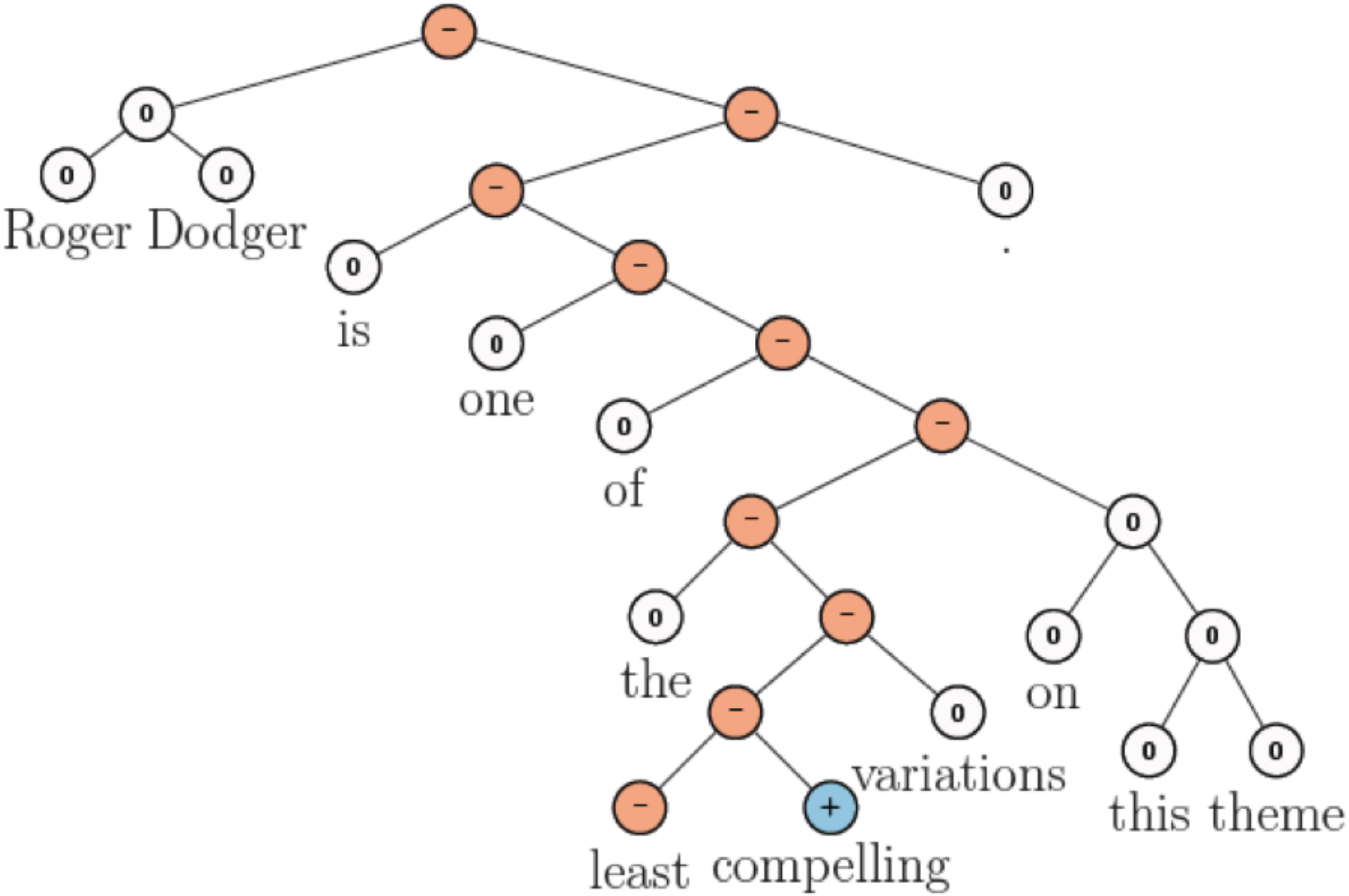
RNTN for Sentiment Analysis



Roger Dodger is one of the most compelling variations on this theme.

Source: Richard Socher et al. (2013) "Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank", EMNLP 2013

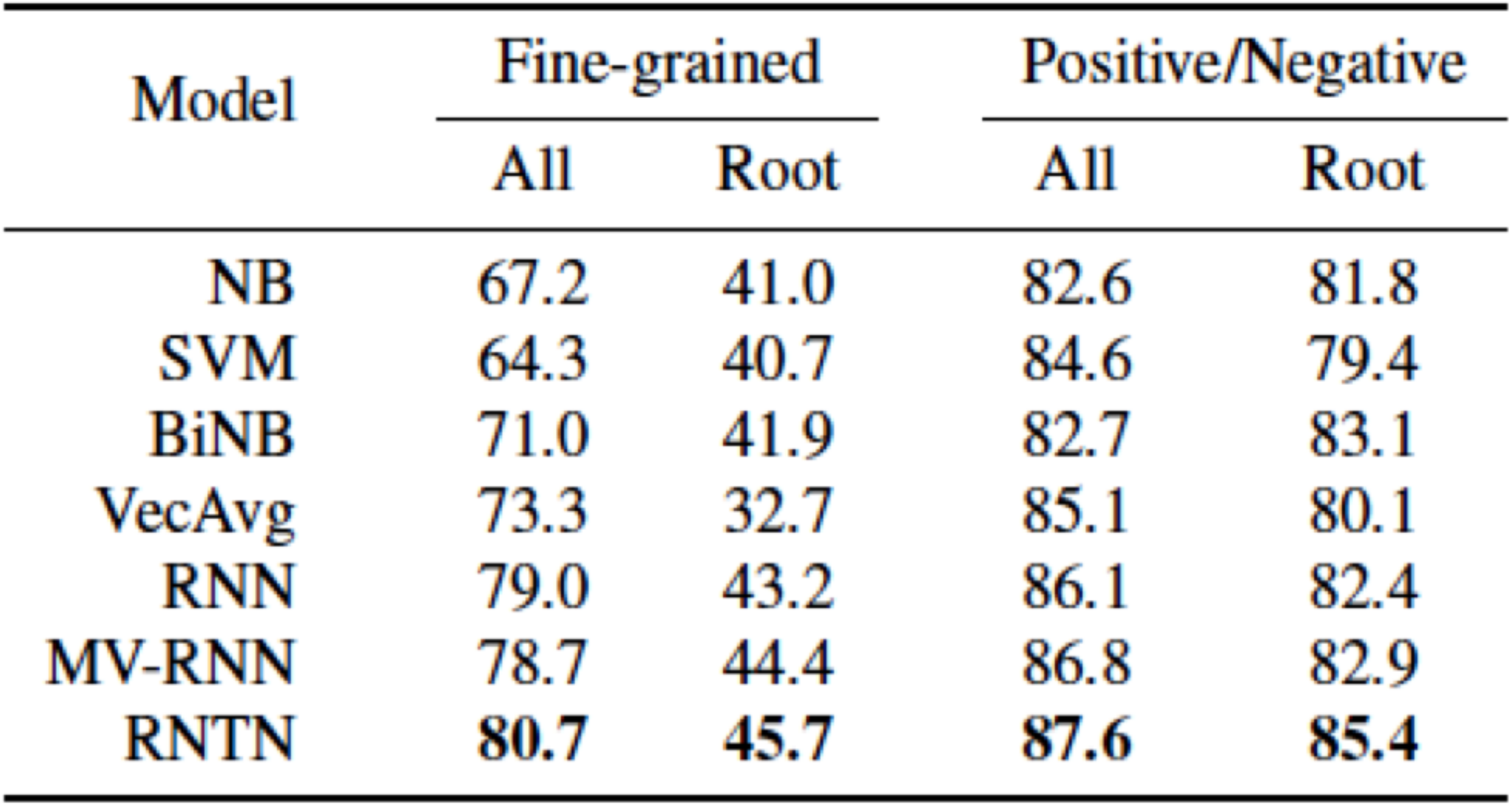
RNTN for Sentiment Analysis



Roger Dodger is one of the least compelling variations on this theme.

Source: Richard Socher et al. (2013) "Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank", EMNLP 2013

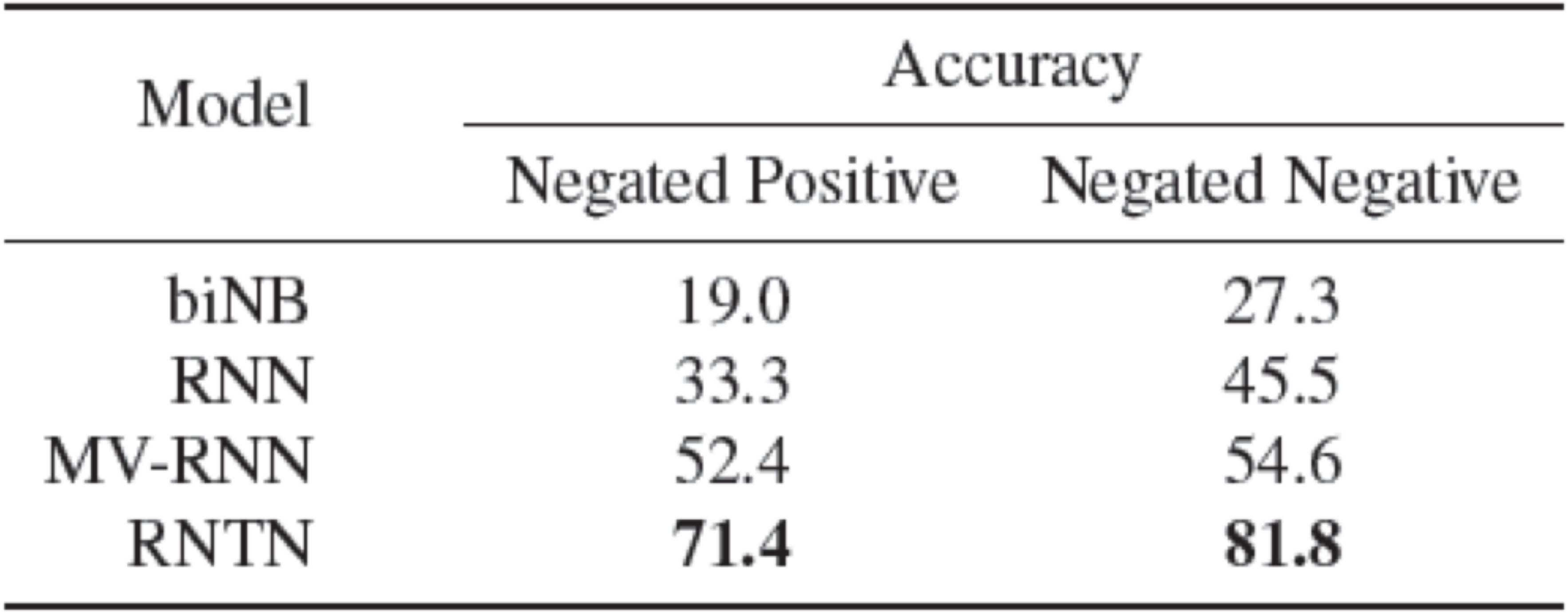
Accuracy for fine grained (5-class) and binary predictions at the sentence level (root) and for all nodes



Model	Fine-grained		Positive/Negative	
	All	Root	All	Root
NB	67.2	41.0	82.6	81.8
SVM	64.3	40.7	84.6	79.4
BiNB	71.0	41.9	82.7	83.1
VecAvg	73.3	32.7	85.1	80.1
RNN	79.0	43.2	86.1	82.4
MV-RNN	78.7	44.4	86.8	82.9
RNTN	**80.7**	**45.7**	**87.6**	**85.4**

Source: Richard Socher et al. (2013) "Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank", EMNLP 2013

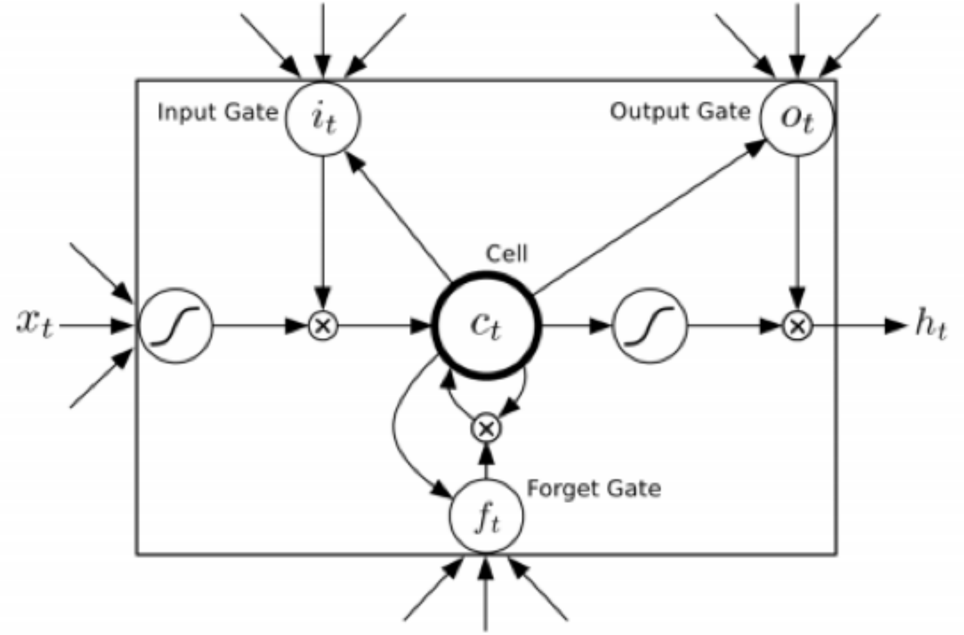
Accuracy of negation detection

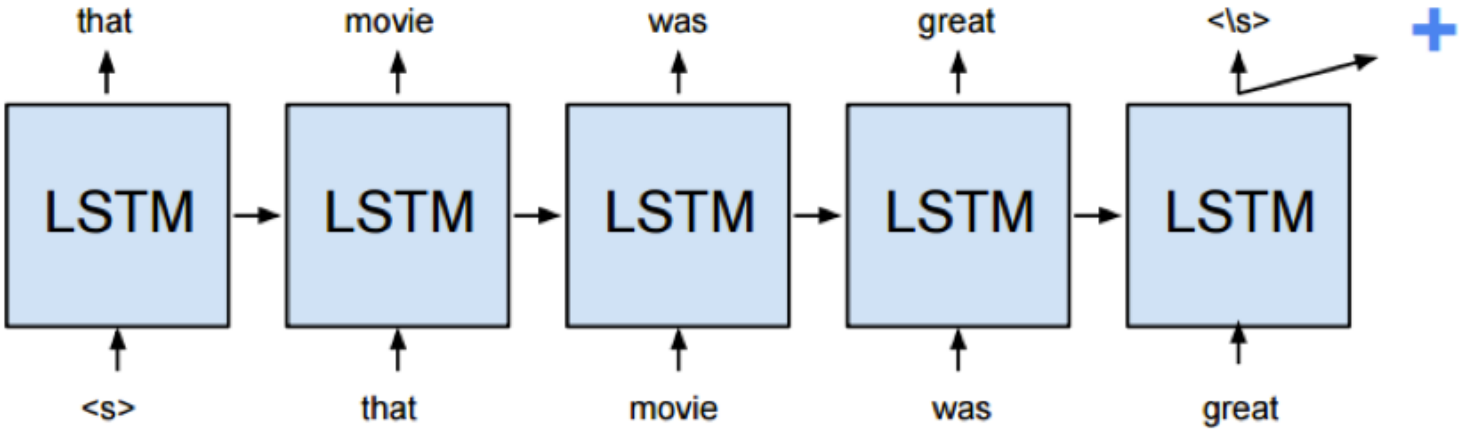


Model	Accuracy	
	Negated Positive	Negated Negative
biNB	19.0	27.3
RNN	33.3	45.5
MV-RNN	52.4	54.6
RNTN	**71.4**	**81.8**

Source: Richard Socher et al. (2013) "Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank", EMNLP 2013

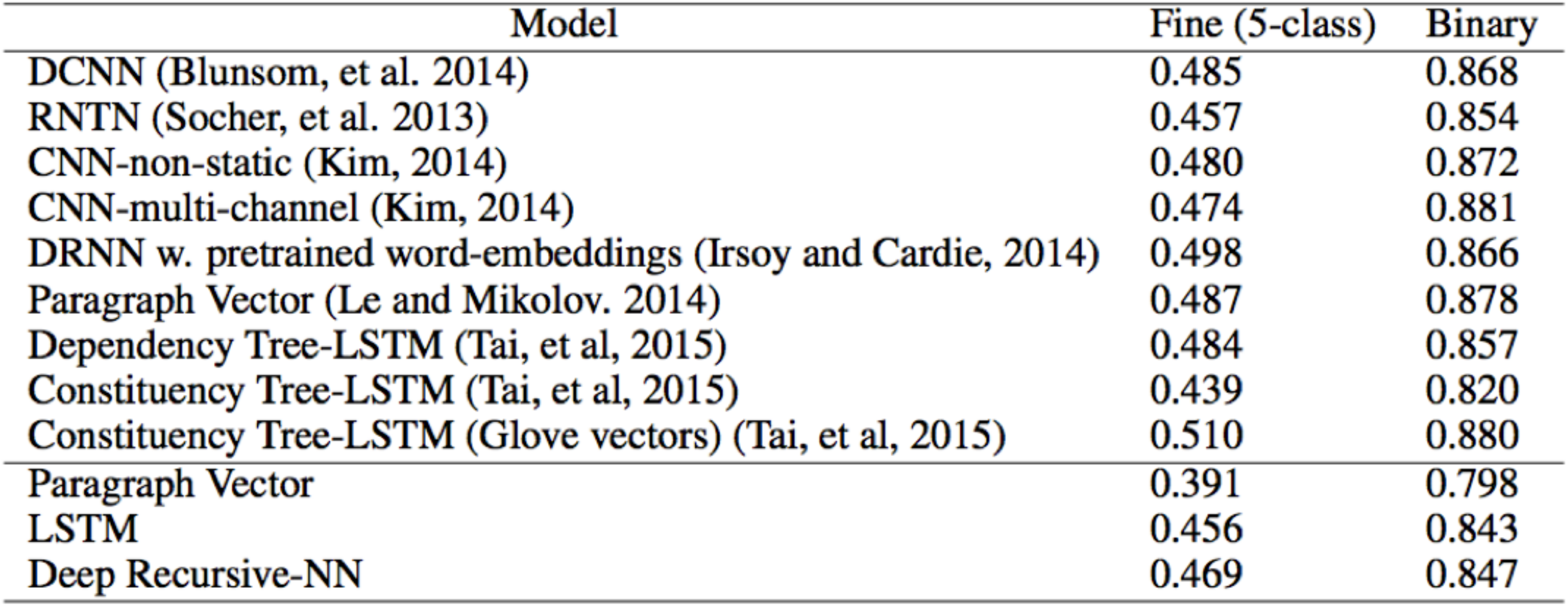
Long Short-Term Memory (LSTM)





Source: https://cs224d.stanford.edu/reports/HongJames.pdf

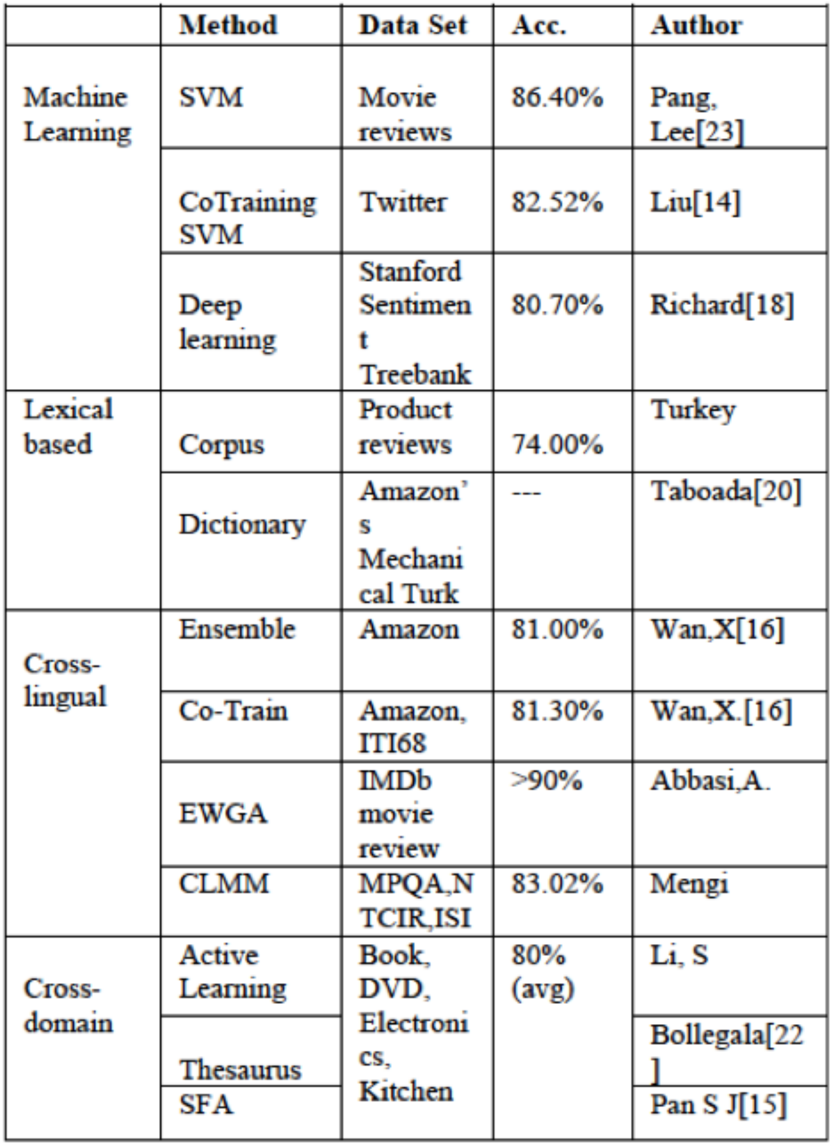
Deep Learning for Sentiment Analysis CNN RNTN LSTM



Model	Fine (5-class)	Binary
DCNN (Blunsom, et al. 2014)	0.485	0.868
RNTN (Socher, et al. 2013)	0.457	0.854
CNN-non-static (Kim, 2014)	0.480	0.872
CNN-multi-channel (Kim, 2014)	0.474	0.881
DRNN w. pretrained word-embeddings (Irsoy and Cardie, 2014)	0.498	0.866
Paragraph Vector (Le and Mikolov. 2014)	0.487	0.878
Dependency Tree-LSTM (Tai, et al, 2015)	0.484	0.857
Constituency Tree-LSTM (Tai, et al, 2015)	0.439	0.820
Constituency Tree-LSTM (Glove vectors) (Tai, et al, 2015)	0.510	0.880
Paragraph Vector	0.391	0.798
LSTM	0.456	0.843
Deep Recursive-NN	0.469	0.847

Source: https://cs224d.stanford.edu/reports/HongJames.pdf

Performance Comparison of Sentiment Analysis Methods



	Method	Data Set	Acc.	Author
Machine Learning	SVM	Movie reviews	86.40%	Pang, Lee[23]
	CoTraining SVM	Twitter	82.52%	Liu[14]
	Deep learning	Stanford Sentiment Treebank	80.70%	Richard[18]
Lexical based	Corpus	Product reviews	74.00%	Turkey
	Dictionary	Amazon’s Mechanical Turk	---	Taboada[20]
Cross-lingual	Ensemble	Amazon	81.00%	Wan,X[16]
	Co-Train	Amazon, ITI68	81.30%	Wan,X.[16]
	EWGA	IMDb movie review	>90%	Abbasi,A.
	CLMM	MPQA,NTCIR,ISI	83.02%	Mengi
Cross-domain	Active Learning	Book, DVD, Electronics, Kitchen	80% (avg)	Li, S
	Thesaurus			Bollegala[22]
	SFA			Pan S J[15]

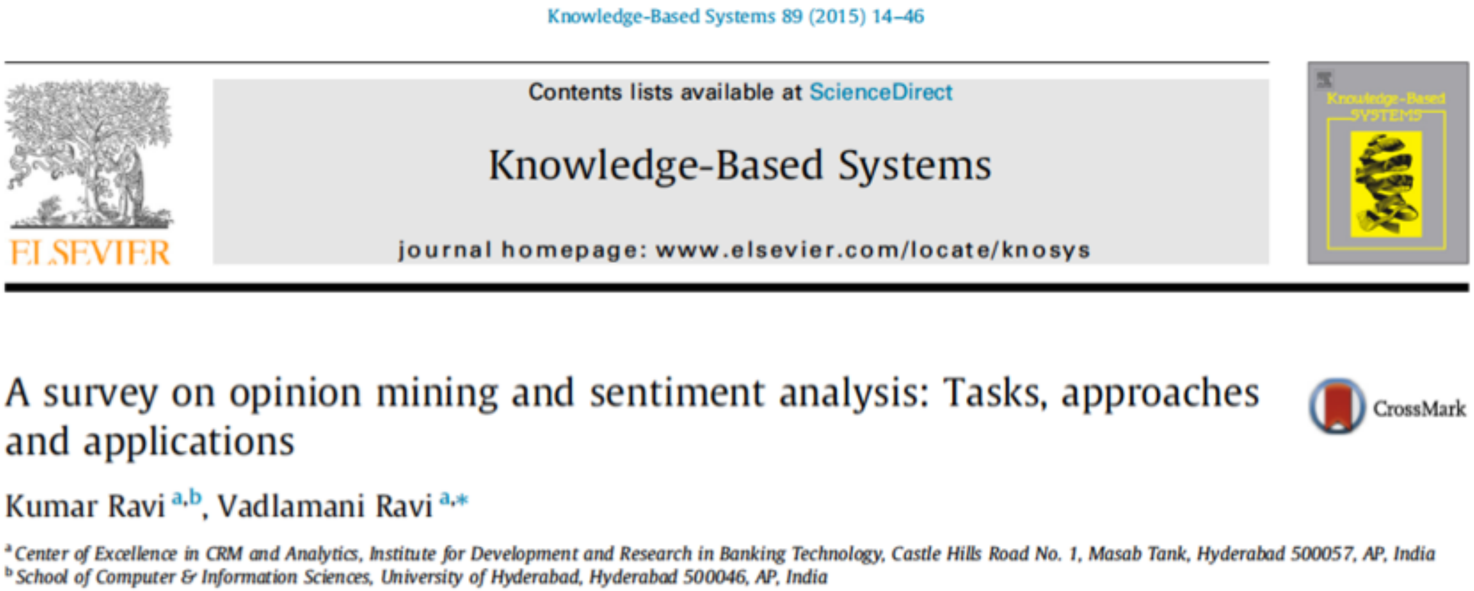
Vishal Kharde and Sheetal Sonawane (2016), "Sentiment Analysis of Twitter Data: A Survey of Techniques," International Journal of Computer Applications, Vol 139, No. 11, 2016. pp.5-15

Kumar Ravi and Vadlamani Ravi (2015),

"A survey on opinion mining and sentiment analysis: tasks, approaches and applications."

Knowledge-Based Systems,

89, pp.14-46



Knowledge-Based Systems 89 (2015) 14–46

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

A survey on opinion mining and sentiment analysis: Tasks, approaches and applications

Kumar Ravi [a,b], Vadlamani Ravi [a,*]

[a] Center of Excellence in CRM and Analytics, Institute for Development and Research in Banking Technology, Castle Hills Road No. 1, Masab Tank, Hyderabad 500057, AP, India
[b] School of Computer & Information Sciences, University of Hyderabad, Hyderabad 500046, AP, India

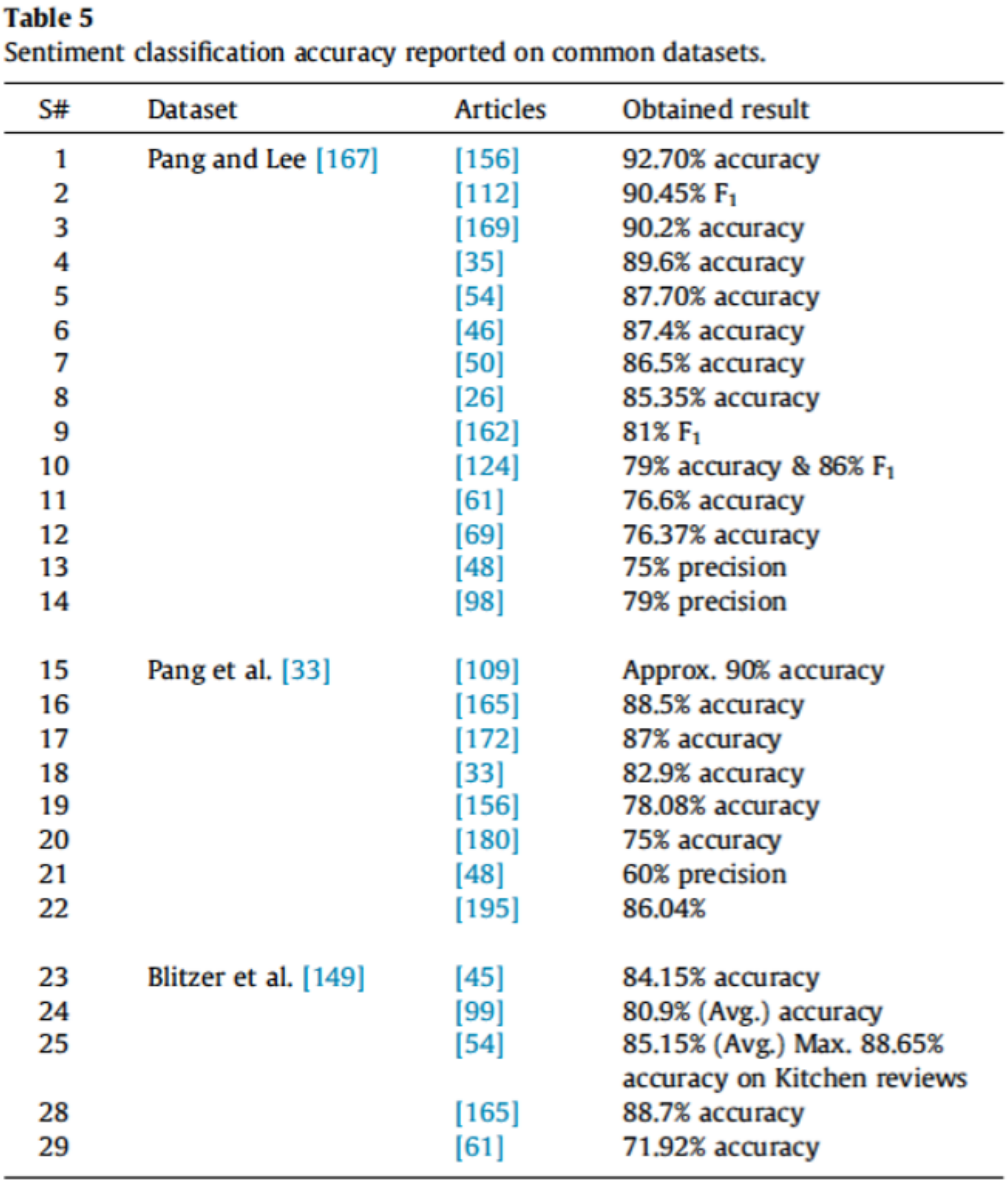
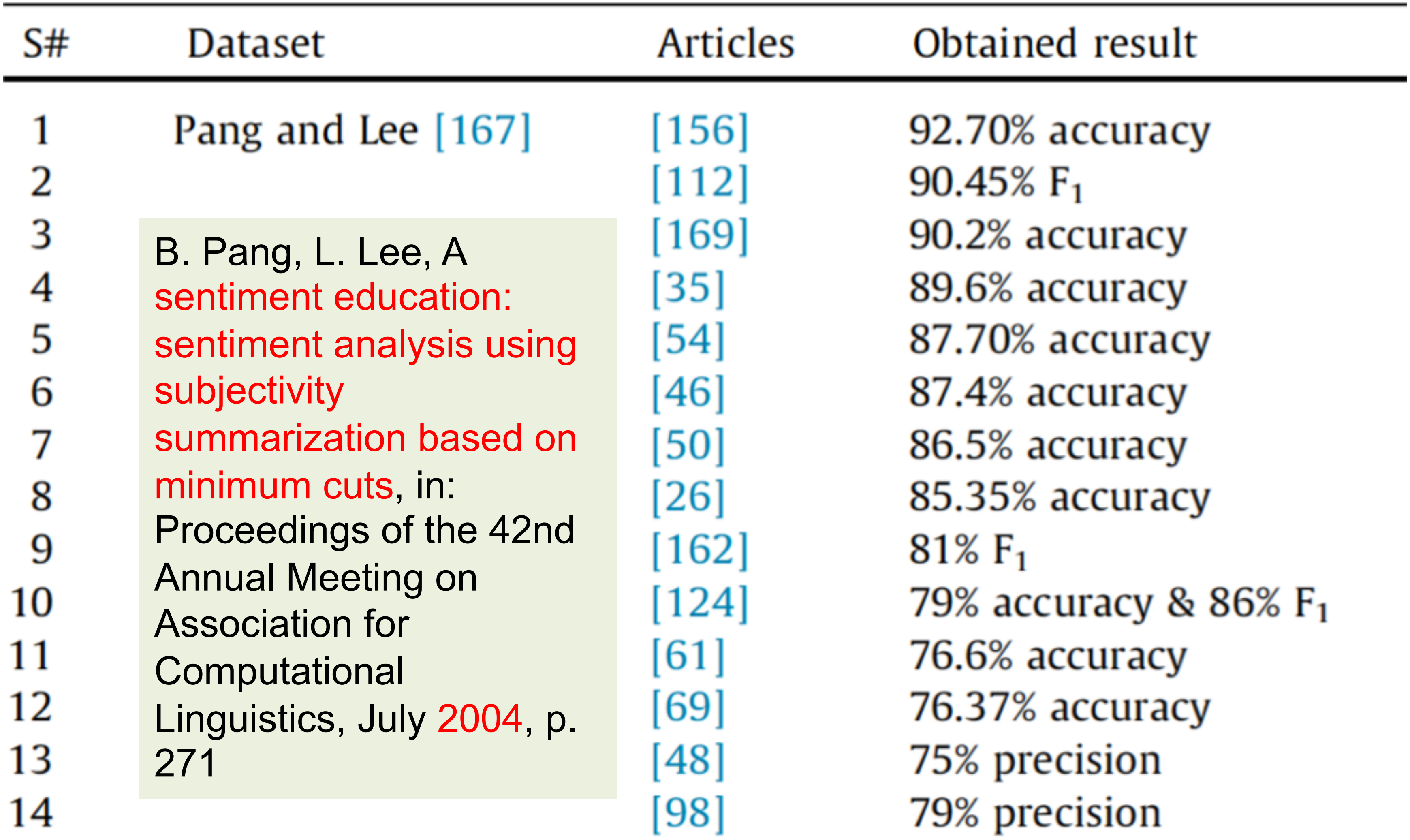


Table 5
Sentiment classification accuracy reported on common datasets.

S#	Dataset	Articles	Obtained result
1	Pang and Lee [167]	[156]	92.70% accuracy
2		[112]	90.45% F_1
3		[169]	90.2% accuracy
4		[35]	89.6% accuracy
5		[54]	87.70% accuracy
6		[46]	87.4% accuracy
7		[50]	86.5% accuracy
8		[26]	85.35% accuracy
9		[162]	81% F_1
10		[124]	79% accuracy & 86% F_1
11		[61]	76.6% accuracy
12		[69]	76.37% accuracy
13		[48]	75% precision
14		[98]	79% precision
15	Pang et al. [33]	[109]	Approx. 90% accuracy
16		[165]	88.5% accuracy
17		[172]	87% accuracy
18		[33]	82.9% accuracy
19		[156]	78.08% accuracy
20		[180]	75% accuracy
21		[48]	60% precision
22		[195]	86.04%
23	Blitzer et al. [149]	[45]	84.15% accuracy
24		[99]	80.9% (Avg.) accuracy
25		[54]	85.15% (Avg.) Max. 88.65% accuracy on Kitchen reviews
28		[165]	88.7% accuracy
29		[61]	71.92% accuracy

Source: Kumar Ravi and Vadlamani Ravi (2015), "A survey on opinion mining and sentiment analysis: tasks, approaches and applications." Knowledge-Based Systems, 89, pp.14-46.

Sentiment Classification Accuracy

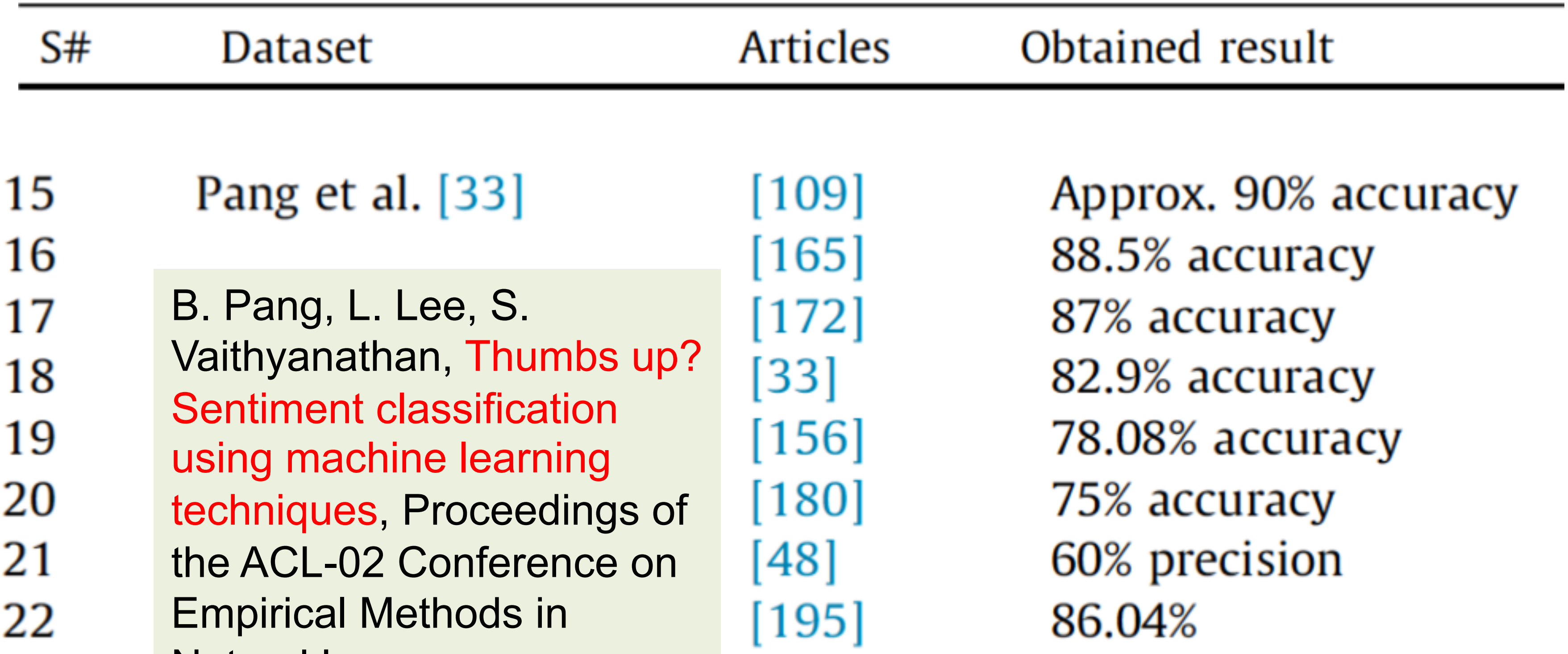


S#	Dataset	Articles	Obtained result
1	Pang and Lee [167]	[156]	92.70% accuracy
2		[112]	90.45% F_1
3		[169]	90.2% accuracy
4		[35]	89.6% accuracy
5		[54]	87.70% accuracy
6		[46]	87.4% accuracy
7		[50]	86.5% accuracy
8		[26]	85.35% accuracy
9		[162]	81% F_1
10		[124]	79% accuracy & 86% F_1
11		[61]	76.6% accuracy
12		[69]	76.37% accuracy
13		[48]	75% precision
14		[98]	79% precision

B. Pang, L. Lee, A sentiment education: sentiment analysis using subjectivity summarization based on minimum cuts, in: Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics, July 2004, p. 271

Source: Kumar Ravi and Vadlamani Ravi (2015), "A survey on opinion mining and sentiment analysis: tasks, approaches and applications." Knowledge-Based Systems, 89, pp.14-46.

Sentiment Classification Accuracy

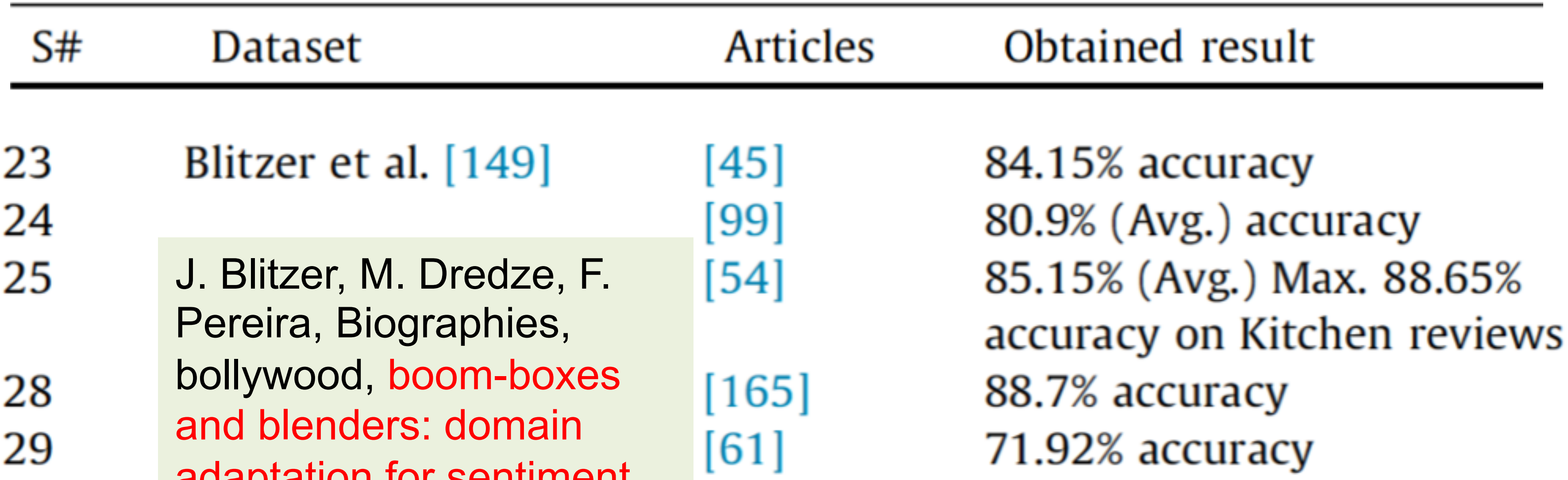


S#	Dataset	Articles	Obtained result
15	Pang et al. [33]	[109]	Approx. 90% accuracy
16		[165]	88.5% accuracy
17		[172]	87% accuracy
18		[33]	82.9% accuracy
19		[156]	78.08% accuracy
20		[180]	75% accuracy
21		[48]	60% precision
22		[195]	86.04%

B. Pang, L. Lee, S. Vaithyanathan, Thumbs up? Sentiment classification using machine learning techniques, Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing, vol. 10, Association for Computational Linguistics, 2002, pp. 79–86.

Source: Kumar Ravi and Vadlamani Ravi (2015), "A survey on opinion mining and sentiment analysis: tasks, approaches and applications." Knowledge-Based Systems, 89, pp.14-46.

Sentiment Classification Accuracy

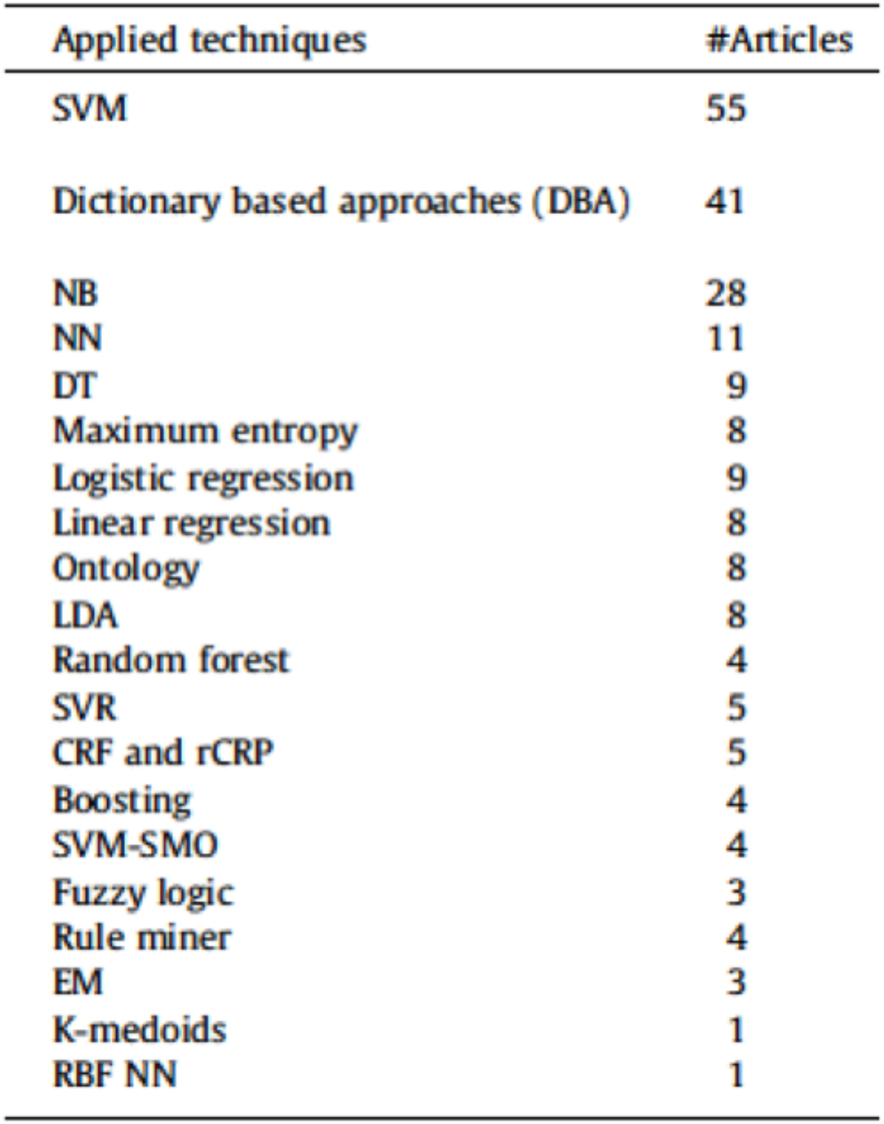


S#	Dataset	Articles	Obtained result
23	Blitzer et al. [149]	[45]	84.15% accuracy
24		[99]	80.9% (Avg.) accuracy
25		[54]	85.15% (Avg.) Max. 88.65% accuracy on Kitchen reviews
28		[165]	88.7% accuracy
29		[61]	71.92% accuracy

J. Blitzer, M. Dredze, F. Pereira, Biographies, bollywood, boom-boxes and blenders: domain adaptation for sentiment classification, in: Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, ACL’07, vol. 7, 2007, pp. 187–205 (13, 29).

Source: Kumar Ravi and Vadlamani Ravi (2015), "A survey on opinion mining and sentiment analysis: tasks, approaches and applications." Knowledge-Based Systems, 89, pp.14-46.

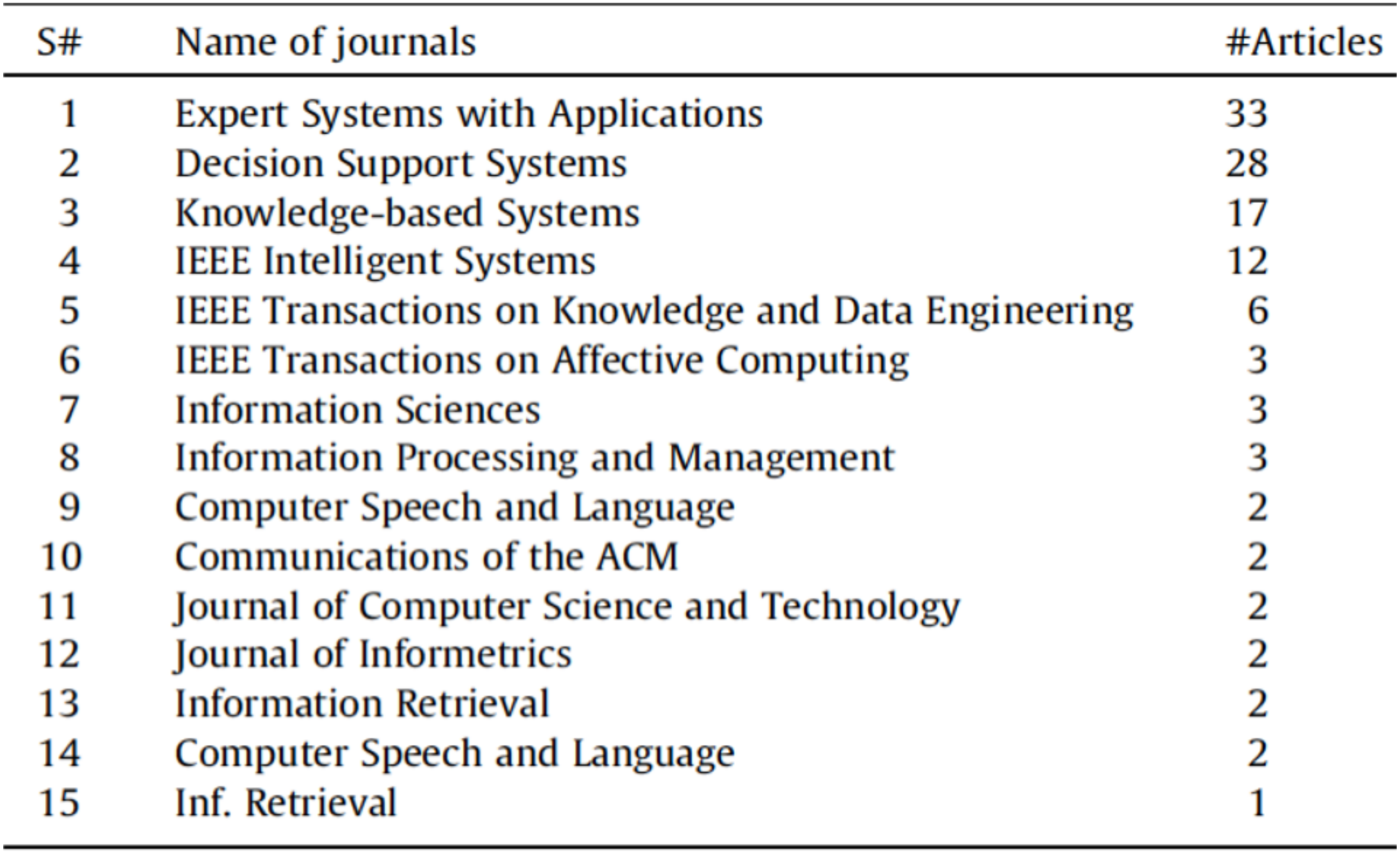
Techniques for Sentiment Analysis



Applied techniques	#Articles
SVM	55
Dictionary based approaches (DBA)	41
NB	28
NN	11
DT	9
Maximum entropy	8
Logistic regression	9
Linear regression	8
Ontology	8
LDA	8
Random forest	4
SVR	5
CRF and rCRP	5
Boosting	4
SVM-SMO	4
Fuzzy logic	3
Rule miner	4
EM	3
K-medoids	1
RBF NN	1

Source: Kumar Ravi and Vadlamani Ravi (2015), "A survey on opinion mining and sentiment analysis: tasks, approaches and applications." Knowledge-Based Systems, 89, pp.14-46.

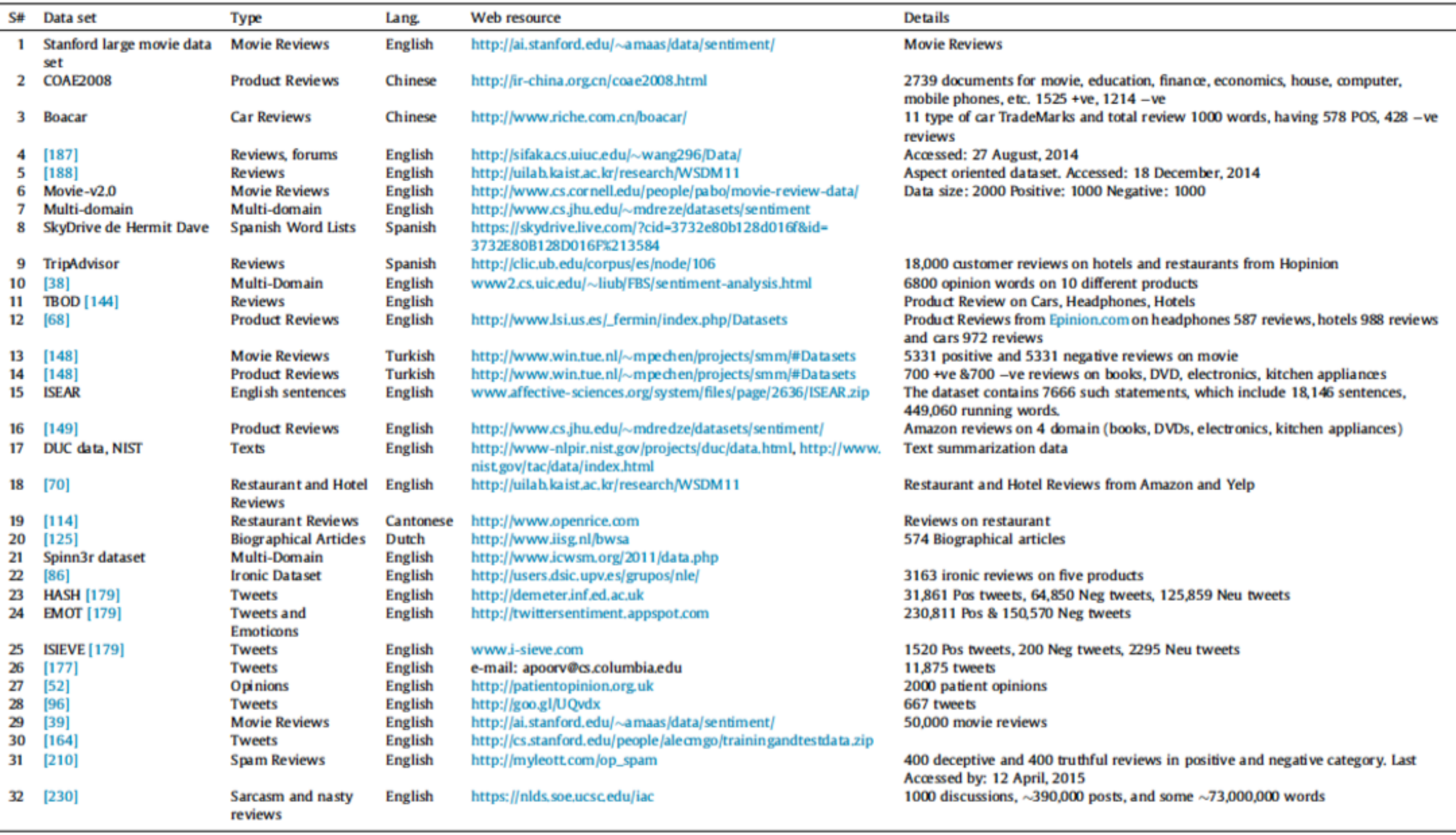
Sentiment Analysis Articles in Journals (2002-2014)



S#	Name of journals	#Articles
1	Expert Systems with Applications	33
2	Decision Support Systems	28
3	Knowledge-based Systems	17
4	IEEE Intelligent Systems	12
5	IEEE Transactions on Knowledge and Data Engineering	6
6	IEEE Transactions on Affective Computing	3
7	Information Sciences	3
8	Information Processing and Management	3
9	Computer Speech and Language	2
10	Communications of the ACM	2
11	Journal of Computer Science and Technology	2
12	Journal of Informetrics	2
13	Information Retrieval	2
14	Computer Speech and Language	2
15	Inf. Retrieval	1

Source: Kumar Ravi and Vadlamani Ravi (2015), "A survey on opinion mining and sentiment analysis: tasks, approaches and applications." Knowledge-Based Systems, 89, pp.14-46.

Publicly Available Datasets for Sentiment Analysis



S#	Data set	Type	Lang.	Web resource	Details
1	Stanford large movie data set	Movie Reviews	English	http://ai.stanford.edu/~amaas/data/sentiment/	Movie Reviews
2	COAE2008	Product Reviews	Chinese	http://ir-china.org.cn/coae2008.html	2739 documents for movie, education, finance, economics, house, computer, mobile phones, etc. 1525 +ve, 1214 –ve
3	Boacar	Car Reviews	Chinese	http://www.riche.com.cn/boacar/	11 type of car TradeMarks and total review 1000 words, having 578 POS, 428 –ve reviews
4	[187]	Reviews, forums	English	http://sifaka.cs.uiuc.edu/~wang296/Data/	Accessed: 27 August, 2014
5	[188]	Reviews	English	http://uilab.kaist.ac.kr/research/WSDM11	Aspect oriented dataset. Accessed: 18 December, 2014
6	Movie-v2.0	Movie Reviews	English	http://www.cs.cornell.edu/people/pabo/movie-review-data/	Data size: 2000 Positive: 1000 Negative: 1000
7	Multi-domain	Multi-domain	English	http://www.cs.jhu.edu/~mdreze/datasets/sentiment	
8	SkyDrive de Hermit Dave	Spanish Word Lists	Spanish	https://skydrive.live.com/?cid=3732e80b128d016f&id=3732E80B128D016F%213584	
9	TripAdvisor	Reviews	Spanish	http://clic.ub.edu/corpus/es/node/106	18,000 customer reviews on hotels and restaurants from Hopinion
10	[38]	Multi-Domain	English	www2.cs.uic.edu/~liub/FBS/sentiment-analysis.html	6800 opinion words on 10 different products
11	TBOD [144]	Reviews	English		Product Review on Cars, Headphones, Hotels
12	[68]	Product Reviews	English	http://www.lsi.us.es/_fermin/index.php/Datasets	Product Reviews from Epinion.com on headphones 587 reviews, hotels 988 reviews and cars 972 reviews
13	[148]	Movie Reviews	Turkish	http://www.win.tue.nl/~mpechen/projects/smm/#Datasets	5331 positive and 5331 negative reviews on movie
14	[148]	Product Reviews	Turkish	http://www.win.tue.nl/~mpechen/projects/smm/#Datasets	700 +ve &700 –ve reviews on books, DVD, electronics, kitchen appliances
15	ISEAR	English sentences	English	www.affective-sciences.org/system/files/page/2636/ISEAR.zip	The dataset contains 7666 such statements, which include 18,146 sentences, 449,060 running words.
16	[149]	Product Reviews	English	http://www.cs.jhu.edu/~mdredze/datasets/sentiment/	Amazon reviews on 4 domain (books, DVDs, electronics, kitchen appliances)
17	DUC data, NIST	Texts	English	http://www-nlpir.nist.gov/projects/duc/data.html, http://www.nist.gov/tac/data/index.html	Text summarization data
18	[70]	Restaurant and Hotel Reviews	English	http://uilab.kaist.ac.kr/research/WSDM11	Restaurant and Hotel Reviews from Amazon and Yelp
19	[114]	Restaurant Reviews	Cantonese	http://www.openrice.com	Reviews on restaurant
20	[125]	Biographical Artides	Dutch	http://www.iisg.nl/bwsa	574 Biographical articles
21	Spinn3r dataset	Multi-Domain	English	http://www.icwsm.org/2011/data.php	
22	[86]	Ironic Dataset	English	http://users.dsic.upv.es/grupos/nle/	3163 ironic reviews on five products
23	HASH [179]	Tweets	English	http://demeter.inf.ed.ac.uk	31,861 Pos tweets, 64,850 Neg tweets, 125,859 Neu tweets
24	EMOT [179]	Tweets and Emoticons	English	http://twittersentiment.appspot.com	230,811 Pos & 150,570 Neg tweets
25	ISIEVE [179]	Tweets	English	www.i-sieve.com	1520 Pos tweets, 200 Neg tweets, 2295 Neu tweets
26	[177]	Tweets	English	e-mail: apoorv@cs.columbia.edu	11,875 tweets
27	[52]	Opinions	English	http://patientopinion.org.uk	2000 patient opinions
28	[96]	Tweets	English	http://goo.gl/UQvdx	667 tweets
29	[39]	Movie Reviews	English	http://ai.stanford.edu/~amaas/data/sentiment/	50,000 movie reviews
30	[164]	Tweets	English	http://cs.stanford.edu/people/alecmgo/trainingandtestdata.zip	
31	[210]	Spam Reviews	English	http://myleott.com/op_spam	400 deceptive and 400 truthful reviews in positive and negative category. Last Accessed by: 12 April, 2015
32	[230]	Sarcasm and nasty reviews	English	https://nlds.soe.ucsc.edu/iac	1000 discussions, ~390,000 posts, and some ~73,000,000 words

Source: Kumar Ravi and Vadlamani Ravi (2015), "A survey on opinion mining and sentiment analysis: tasks, approaches and applications." Knowledge-Based Systems, 89, pp.14-46.

Sentiment Analysis Datasets

- Product Reviews (PR)
- Movie Reviews (MR)
- Restaurant Reviews (RR)
- Micro-blog (MB)
- Global domain (G)

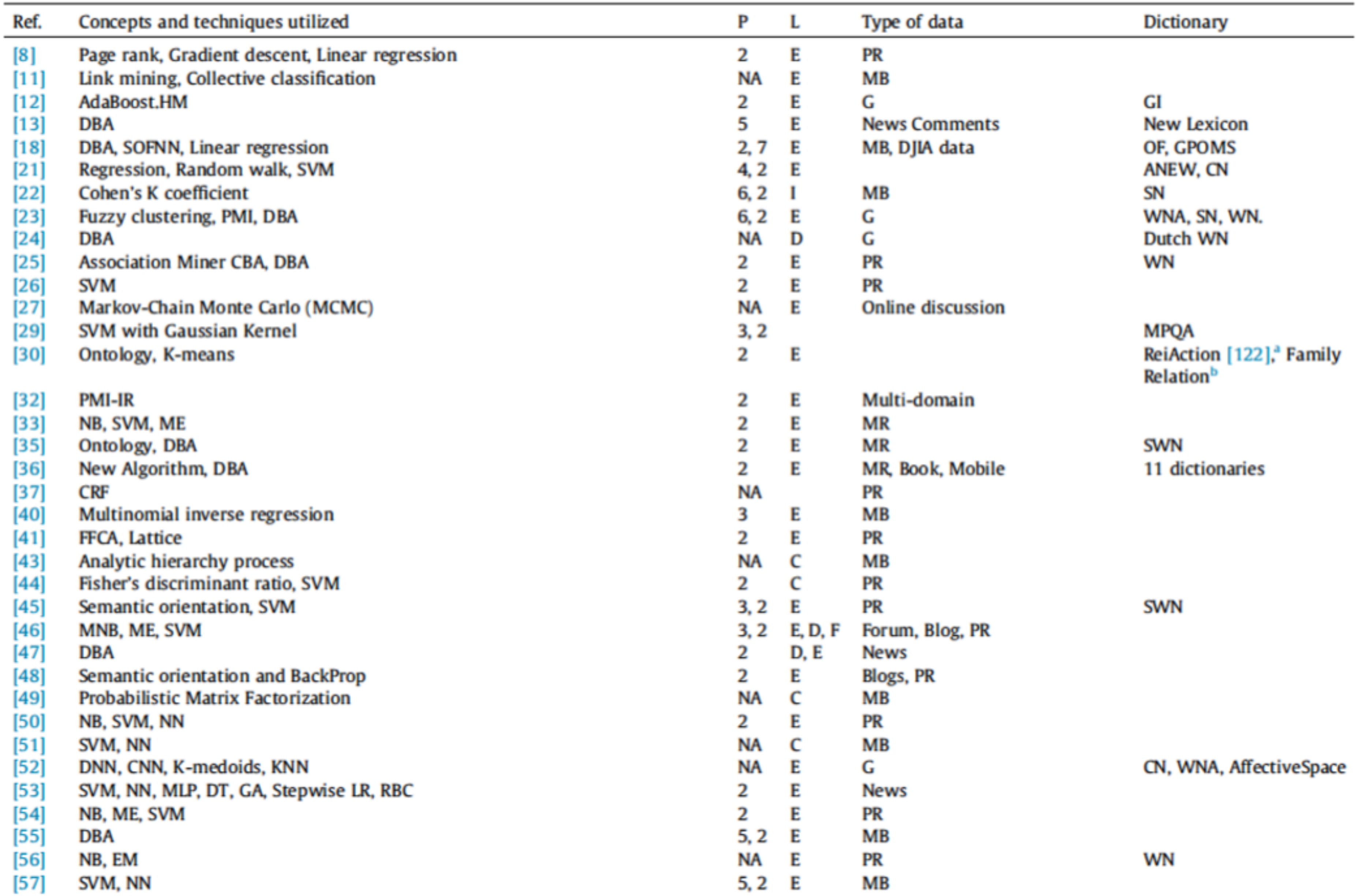
Source: Kumar Ravi and Vadlamani Ravi (2015), "A survey on opinion mining and sentiment analysis: tasks, approaches and applications." Knowledge-Based Systems, 89, pp.14-46.

Sentiment Analysis Dictionary

- SenticNet (SN)
- WordNet (WN)
- ConceptNet (CN)
- WordNet-Affect (WNA)
- Bing Liu Opinion Lexicon (OL)

Source: Kumar Ravi and Vadlamani Ravi (2015), "A survey on opinion mining and sentiment analysis: tasks, approaches and applications." Knowledge-Based Systems, 89, pp.14-46.

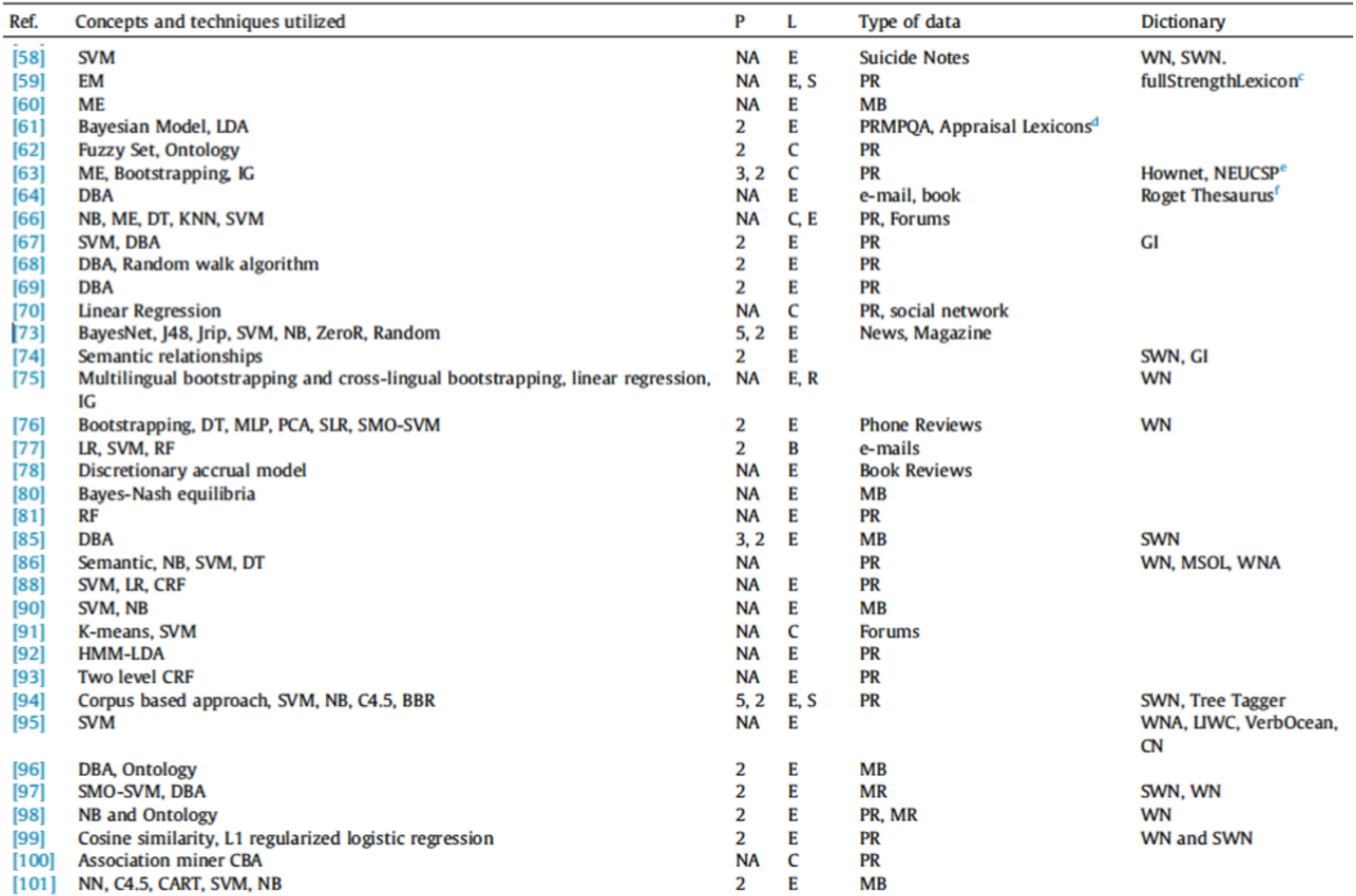
Summary of reviewed articles



Ref.	Concepts and techniques utilized	P	L	Type of data	Dictionary
[8]	Page rank, Gradient descent, Linear regression	2	E	PR	
[11]	Link mining, Collective classification	NA	E	MB	
[12]	AdaBoost.HM	2	E	G	GI
[13]	DBA	5	E	News Comments	New Lexicon
[18]	DBA, SOFNN, Linear regression	2, 7	E	MB, DJIA data	OF, GPOMS
[21]	Regression, Random walk, SVM	4, 2	E		ANEW, CN
[22]	Cohen's K coefficient	6, 2	I	MB	SN
[23]	Fuzzy clustering, PMI, DBA	6, 2	E	G	WNA, SN, WN.
[24]	DBA	NA	D	G	Dutch WN
[25]	Association Miner CBA, DBA	2	E	PR	WN
[26]	SVM	2	E	PR	
[27]	Markov-Chain Monte Carlo (MCMC)	NA	E	Online discussion	
[29]	SVM with Gaussian Kernel	3, 2			MPQA
[30]	Ontology, K-means	2	E		ReiAction [122],[a] Family Relation[b]
[32]	PMI-IR	2	E	Multi-domain	
[33]	NB, SVM, ME	2	E	MR	
[35]	Ontology, DBA	2	E	MR	SWN
[36]	New Algorithm, DBA	2	E	MR, Book, Mobile	11 dictionaries
[37]	CRF	NA		PR	
[40]	Multinomial inverse regression	3	E	MB	
[41]	FFCA, Lattice	2	E	PR	
[43]	Analytic hierarchy process	NA	C	MB	
[44]	Fisher's discriminant ratio, SVM	2	C	PR	
[45]	Semantic orientation, SVM	3, 2	E	PR	SWN
[46]	MNB, ME, SVM	3, 2	E, D, F	Forum, Blog, PR	
[47]	DBA	2	D, E	News	
[48]	Semantic orientation and BackProp	2	E	Blogs, PR	
[49]	Probabilistic Matrix Factorization	NA	C	MB	
[50]	NB, SVM, NN	2	E	PR	
[51]	SVM, NN	NA	C	MB	
[52]	DNN, CNN, K-medoids, KNN	NA	E	G	CN, WNA, AffectiveSpace
[53]	SVM, NN, MLP, DT, GA, Stepwise LR, RBC	2	E	News	
[54]	NB, ME, SVM	2	E	PR	
[55]	DBA	5, 2	E	MB	
[56]	NB, EM	NA	E	PR	WN
[57]	SVM, NN	5, 2	E	MB	

Source: Kumar Ravi and Vadlamani Ravi (2015), "A survey on opinion mining and sentiment analysis: tasks, approaches and applications." Knowledge-Based Systems, 89, pp.14-46.

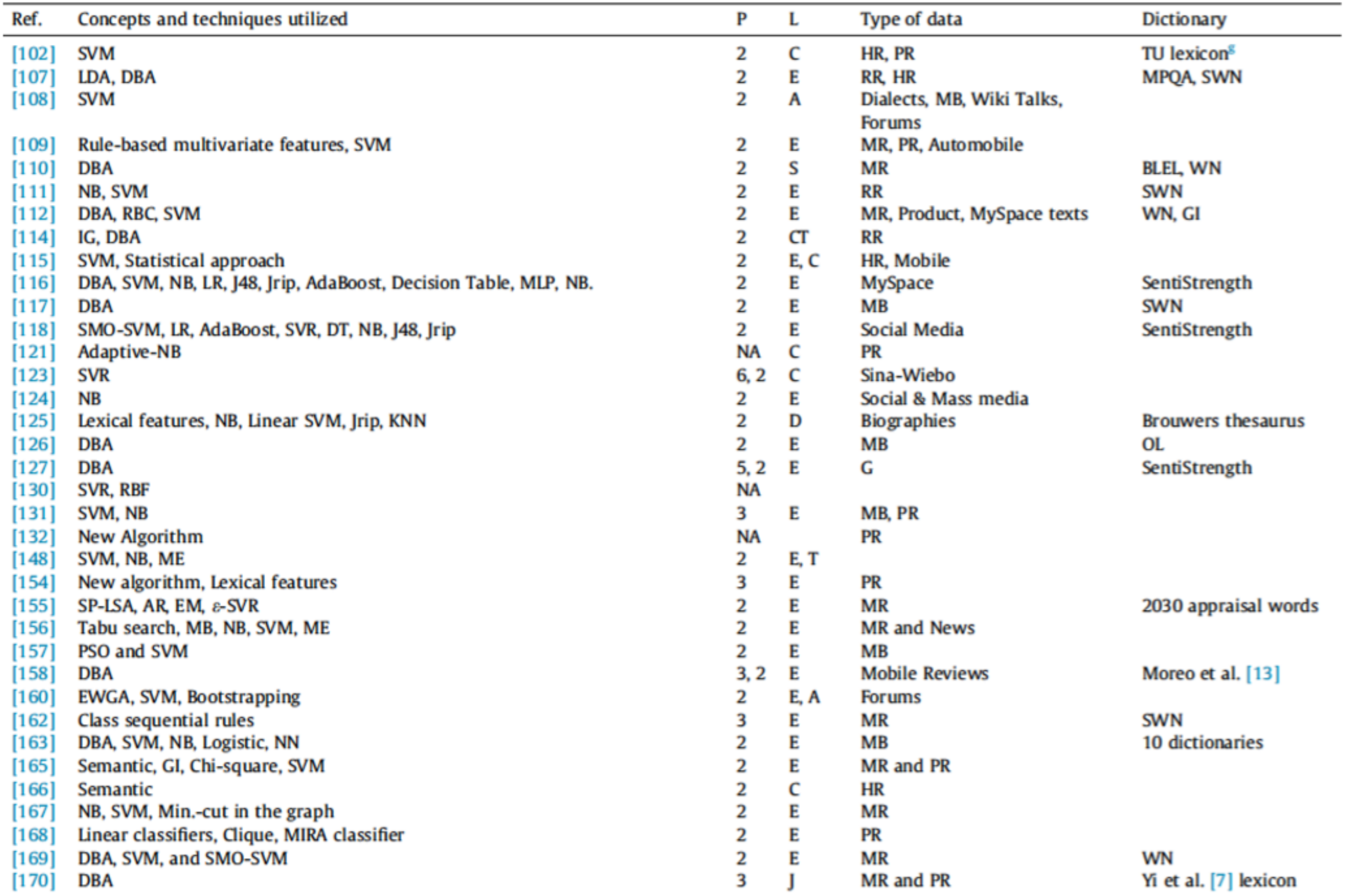
Summary of reviewed articles



Ref.	Concepts and techniques utilized	P	L	Type of data	Dictionary
[58]	SVM	NA	E	Suicide Notes	WN, SWN.
[59]	EM	NA	E, S	PR	fullStrengthLexicon[c]
[60]	ME	NA	E	MB	
[61]	Bayesian Model, LDA	2	E	PRMPQA, Appraisal Lexicons[d]	
[62]	Fuzzy Set, Ontology	2	C	PR	
[63]	ME, Bootstrapping, IG	3, 2	C	PR	Hownet, NEUCSP[e]
[64]	DBA	NA	E	e-mail, book	Roget Thesaurus[f]
[66]	NB, ME, DT, KNN, SVM	NA	C, E	PR, Forums	
[67]	SVM, DBA	2	E	PR	GI
[68]	DBA, Random walk algorithm	2	E	PR	
[69]	DBA	2	E	PR	
[70]	Linear Regression	NA	C	PR, social network	
[73]	BayesNet, J48, Jrip, SVM, NB, ZeroR, Random	5, 2	E	News, Magazine	
[74]	Semantic relationships	2	E		SWN, GI
[75]	Multilingual bootstrapping and cross-lingual bootstrapping, linear regression, IG	NA	E, R		WN
[76]	Bootstrapping, DT, MLP, PCA, SLR, SMO-SVM	2	E	Phone Reviews	WN
[77]	LR, SVM, RF	2	B	e-mails	
[78]	Discretionary accrual model	NA	E	Book Reviews	
[80]	Bayes-Nash equilibria	NA	E	MB	
[81]	RF	NA	E	PR	
[85]	DBA	3, 2	E	MB	SWN
[86]	Semantic, NB, SVM, DT	NA		PR	WN, MSOL, WNA
[88]	SVM, LR, CRF	NA	E	PR	
[90]	SVM, NB	NA	E	MB	
[91]	K-means, SVM	NA	C	Forums	
[92]	HMM-LDA	NA	E	PR	
[93]	Two level CRF	NA	E	PR	
[94]	Corpus based approach, SVM, NB, C4.5, BBR	5, 2	E, S	PR	SWN, Tree Tagger
[95]	SVM	NA	E		WNA, LIWC, VerbOcean, CN
[96]	DBA, Ontology	2	E	MB	
[97]	SMO-SVM, DBA	2	E	MR	SWN, WN
[98]	NB and Ontology	2	E	PR, MR	WN
[99]	Cosine similarity, L1 regularized logistic regression	2	E	PR	WN and SWN
[100]	Association miner CBA	NA	C	PR	
[101]	NN, C4.5, CART, SVM, NB	2	E	MB	

Source: Kumar Ravi and Vadlamani Ravi (2015), "A survey on opinion mining and sentiment analysis: tasks, approaches and applications." Knowledge-Based Systems, 89, pp.14-46.

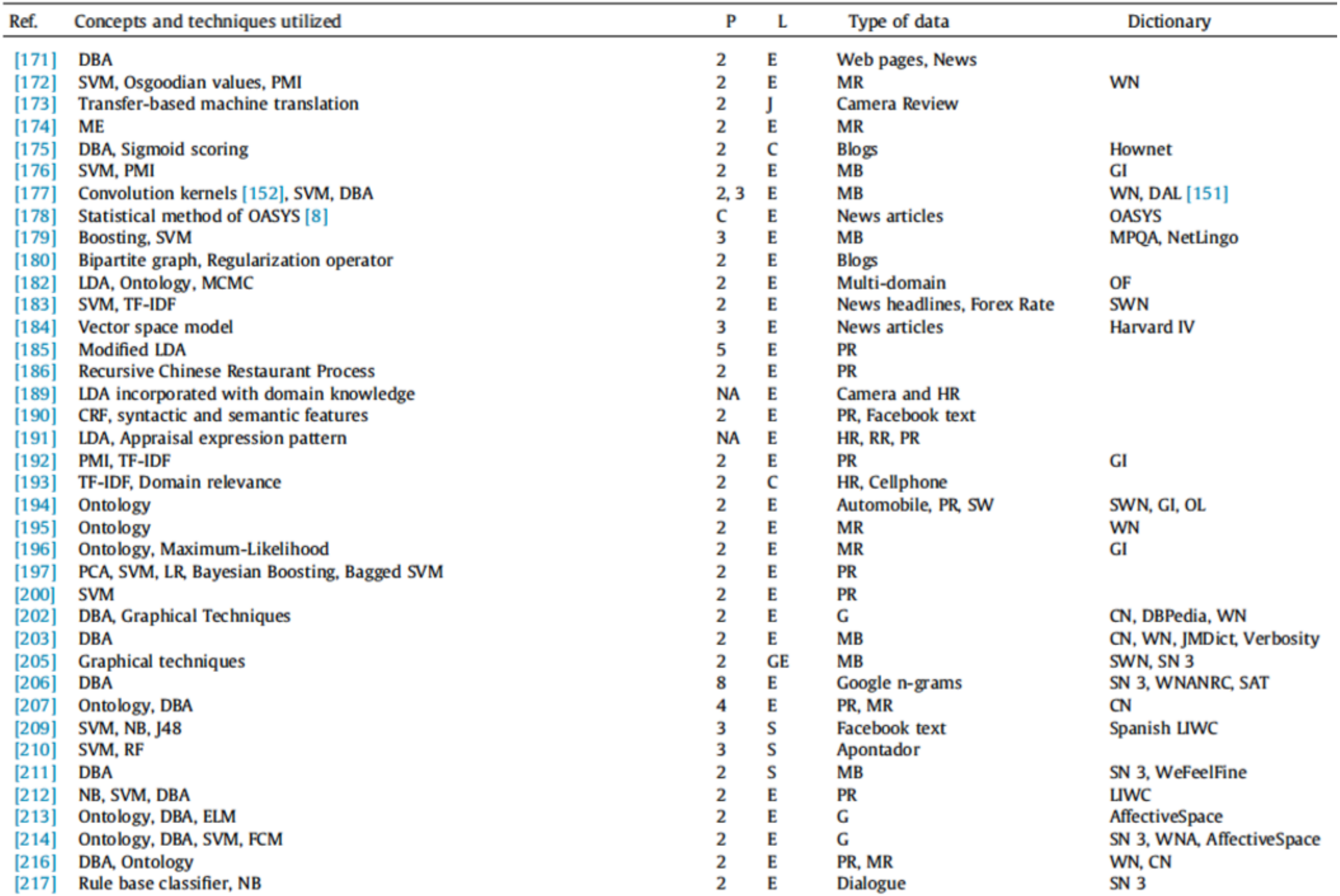
Summary of reviewed articles



Ref.	Concepts and techniques utilized	P	L	Type of data	Dictionary
[102]	SVM	2	C	HR, PR	TU lexicon[g]
[107]	LDA, DBA	2	E	RR, HR	MPQA, SWN
[108]	SVM	2	A	Dialects, MB, Wiki Talks, Forums	
[109]	Rule-based multivariate features, SVM	2	E	MR, PR, Automobile	
[110]	DBA	2	S	MR	BLEL, WN
[111]	NB, SVM	2	E	RR	SWN
[112]	DBA, RBC, SVM	2	E	MR, Product, MySpace texts	WN, GI
[114]	IG, DBA	2	CT	RR	
[115]	SVM, Statistical approach	2	E, C	HR, Mobile	
[116]	DBA, SVM, NB, LR, J48, Jrip, AdaBoost, Decision Table, MLP, NB.	2	E	MySpace	SentiStrength
[117]	DBA	2	E	MB	SWN
[118]	SMO-SVM, LR, AdaBoost, SVR, DT, NB, J48, Jrip	2	E	Social Media	SentiStrength
[121]	Adaptive-NB	NA	C	PR	
[123]	SVR	6, 2	C	Sina-Wiebo	
[124]	NB	2	E	Social & Mass media	
[125]	Lexical features, NB, Linear SVM, Jrip, KNN	2	D	Biographies	Brouwers thesaurus
[126]	DBA	2	E	MB	OL
[127]	DBA	5, 2	E	G	SentiStrength
[130]	SVR, RBF	NA			
[131]	SVM, NB	3	E	MB, PR	
[132]	New Algorithm	NA		PR	
[148]	SVM, NB, ME	2	E, T		
[154]	New algorithm, Lexical features	3	E	PR	
[155]	SP-LSA, AR, EM, ε-SVR	2	E	MR	2030 appraisal words
[156]	Tabu search, MB, NB, SVM, ME	2	E	MR and News	
[157]	PSO and SVM	2	E	MB	
[158]	DBA	3, 2	E	Mobile Reviews	Moreo et al. [13]
[160]	EWGA, SVM, Bootstrapping	2	E, A	Forums	
[162]	Class sequential rules	3	E	MR	SWN
[163]	DBA, SVM, NB, Logistic, NN	2	E	MB	10 dictionaries
[165]	Semantic, GI, Chi-square, SVM	2	E	MR and PR	
[166]	Semantic	2	C	HR	
[167]	NB, SVM, Min.-cut in the graph	2	E	MR	
[168]	Linear classifiers, Clique, MIRA classifier	2	E	PR	
[169]	DBA, SVM, and SMO-SVM	2	E	MR	WN
[170]	DBA	3	J	MR and PR	Yi et al. [7] lexicon

Source: Kumar Ravi and Vadlamani Ravi (2015), "A survey on opinion mining and sentiment analysis: tasks, approaches and applications." Knowledge-Based Systems, 89, pp.14-46.

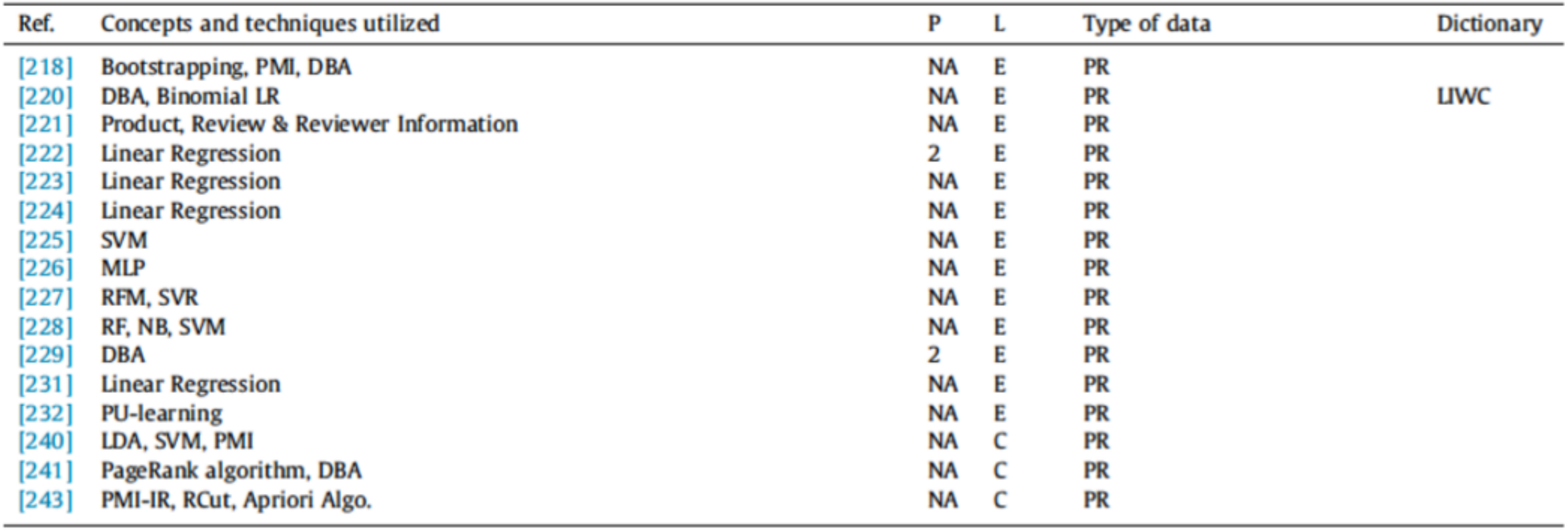
Summary of reviewed articles



Ref.	Concepts and techniques utilized	P	L	Type of data	Dictionary
[171]	DBA	2	E	Web pages, News	
[172]	SVM, Osgoodian values, PMI	2	E	MR	WN
[173]	Transfer-based machine translation	2	J	Camera Review	
[174]	ME	2	E	MR	
[175]	DBA, Sigmoid scoring	2	C	Blogs	Hownet
[176]	SVM, PMI	2	E	MB	GI
[177]	Convolution kernels [152], SVM, DBA	2, 3	E	MB	WN, DAL [151]
[178]	Statistical method of OASYS [8]	C	E	News articles	OASYS
[179]	Boosting, SVM	3	E	MB	MPQA, NetLingo
[180]	Bipartite graph, Regularization operator	2	E	Blogs	
[182]	LDA, Ontology, MCMC	2	E	Multi-domain	OF
[183]	SVM, TF-IDF	2	E	News headlines, Forex Rate	SWN
[184]	Vector space model	3	E	News articles	Harvard IV
[185]	Modified LDA	5	E	PR	
[186]	Recursive Chinese Restaurant Process	2	E	PR	
[189]	LDA incorporated with domain knowledge	NA	E	Camera and HR	
[190]	CRF, syntactic and semantic features	2	E	PR, Facebook text	
[191]	LDA, Appraisal expression pattern	NA	E	HR, RR, PR	
[192]	PMI, TF-IDF	2	E	PR	GI
[193]	TF-IDF, Domain relevance	2	C	HR, Cellphone	
[194]	Ontology	2	E	Automobile, PR, SW	SWN, GI, OL
[195]	Ontology	2	E	MR	WN
[196]	Ontology, Maximum-Likelihood	2	E	MR	GI
[197]	PCA, SVM, LR, Bayesian Boosting, Bagged SVM	2	E	PR	
[200]	SVM	2	E	PR	
[202]	DBA, Graphical Techniques	2	E	G	CN, DBPedia, WN
[203]	DBA	2	E	MB	CN, WN, JMDict, Verbosity
[205]	Graphical techniques	2	GE	MB	SWN, SN 3
[206]	DBA	8	E	Google n-grams	SN 3, WNANRC, SAT
[207]	Ontology, DBA	4	E	PR, MR	CN
[209]	SVM, NB, J48	3	S	Facebook text	Spanish LIWC
[210]	SVM, RF	3	S	Apontador	
[211]	DBA	2	S	MB	SN 3, WeFeelFine
[212]	NB, SVM, DBA	2	E	PR	LIWC
[213]	Ontology, DBA, ELM	2	E	G	AffectiveSpace
[214]	Ontology, DBA, SVM, FCM	2	E	G	SN 3, WNA, AffectiveSpace
[216]	DBA, Ontology	2	E	PR, MR	WN, CN
[217]	Rule base classifier, NB	2	E	Dialogue	SN 3

Source: Kumar Ravi and Vadlamani Ravi (2015), "A survey on opinion mining and sentiment analysis: tasks, approaches and applications." Knowledge-Based Systems, 89, pp.14-46.

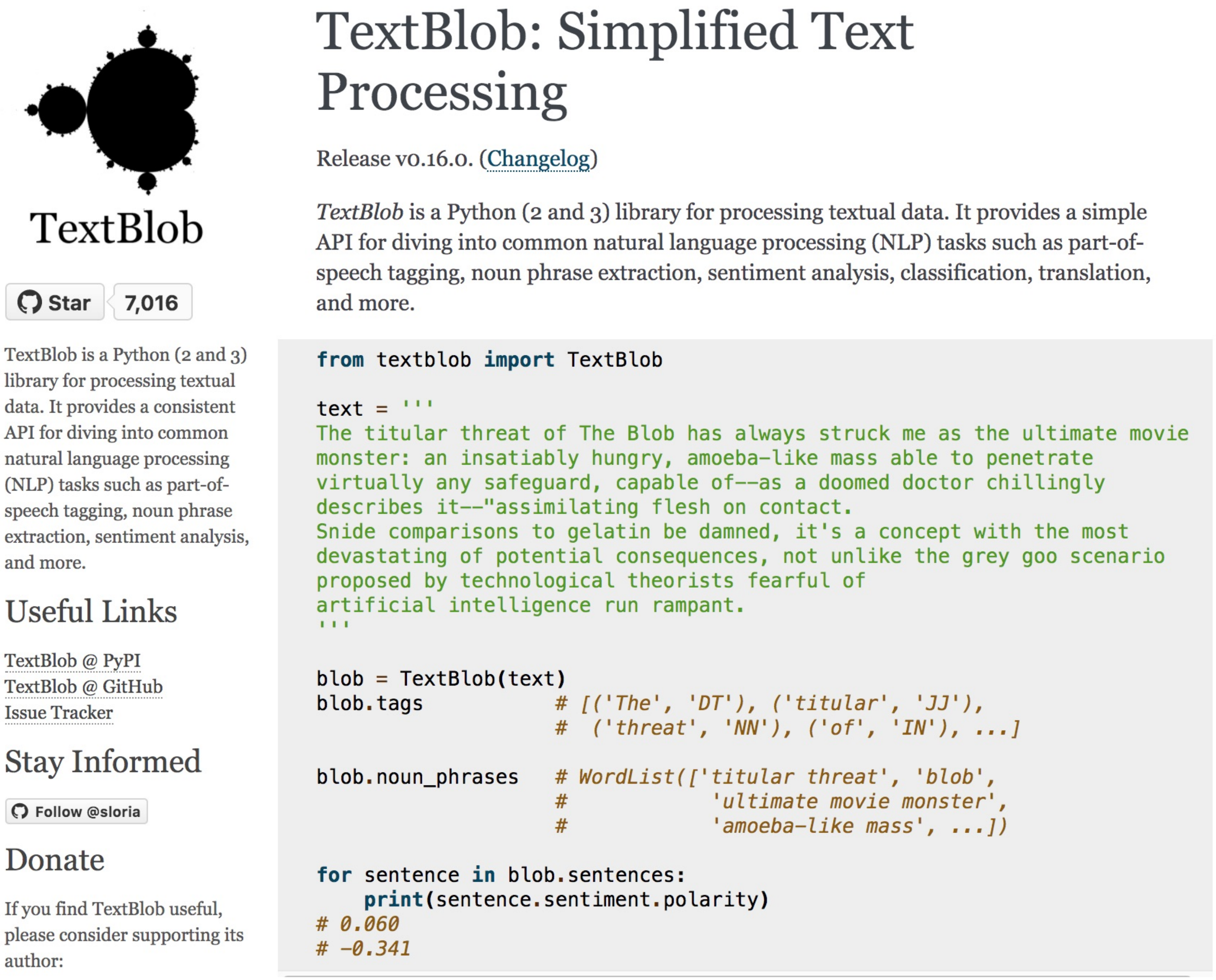
Summary of reviewed articles



Ref.	Concepts and techniques utilized	P	L	Type of data	Dictionary
[218]	Bootstrapping, PMI, DBA	NA	E	PR	
[220]	DBA, Binomial LR	NA	E	PR	LIWC
[221]	Product, Review & Reviewer Information	NA	E	PR	
[222]	Linear Regression	2	E	PR	
[223]	Linear Regression	NA	E	PR	
[224]	Linear Regression	NA	E	PR	
[225]	SVM	NA	E	PR	
[226]	MLP	NA	E	PR	
[227]	RFM, SVR	NA	E	PR	
[228]	RF, NB, SVM	NA	E	PR	
[229]	DBA	2	E	PR	
[231]	Linear Regression	NA	E	PR	
[232]	PU-learning	NA	E	PR	
[240]	LDA, SVM, PMI	NA	C	PR	
[241]	PageRank algorithm, DBA	NA	C	PR	
[243]	PMI-IR, RCut, Apriori Algo.	NA	C	PR	

Source: Kumar Ravi and Vadlamani Ravi (2015), "A survey on opinion mining and sentiment analysis: tasks, approaches and applications." Knowledge-Based Systems, 89, pp.14-46.

TextBlob



TextBlob

Star 7,016

TextBlob is a Python (2 and 3) library for processing textual data. It provides a consistent API for diving into common natural language processing (NLP) tasks such as part-of-speech tagging, noun phrase extraction, sentiment analysis, and more.

Useful Links

TextBlob @ PyPI
TextBlob @ GitHub
Issue Tracker

Stay Informed

Follow @sloria

Donate

If you find TextBlob useful, please consider supporting its author:

TextBlob: Simplified Text Processing

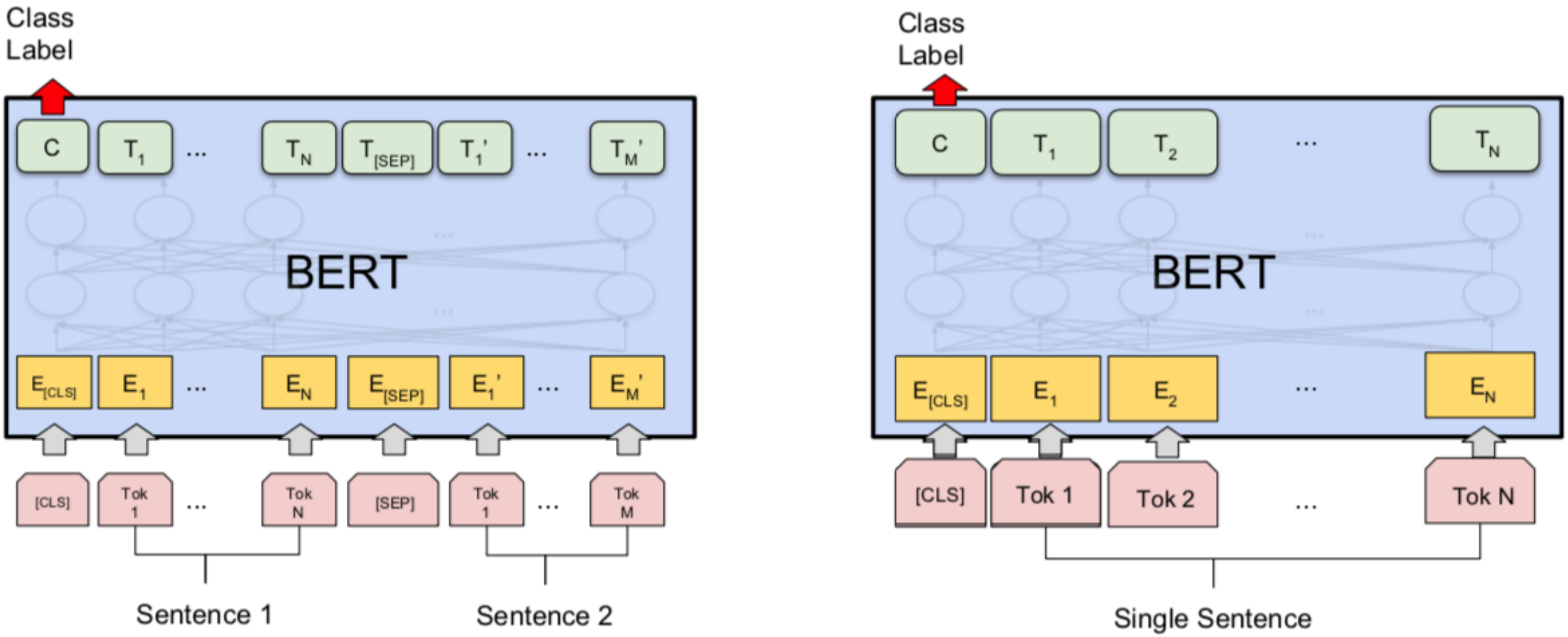
Release v0.16.0. (Changelog)

TextBlob is a Python (2 and 3) library for processing textual data. It provides a simple API for diving into common natural language processing (NLP) tasks such as part-of-speech tagging, noun phrase extraction, sentiment analysis, classification, translation, and more.

```
from textblob import TextBlob

text = '''
The titular threat of The Blob has always struck me as the ultimate movie
monster: an insatiably hungry, amoeba-like mass able to penetrate
virtually any safeguard, capable of--as a doomed doctor chillingly
describes it--"assimilating flesh on contact.
Snide comparisons to gelatin be damned, it's a concept with the most
devastating of potential consequences, not unlike the grey goo scenario
proposed by technological theorists fearful of
artificial intelligence run rampant.
'''

blob = TextBlob(text)
blob.tags           # [('The', 'DT'), ('titular', 'JJ'),
                    #  ('threat', 'NN'), ('of', 'IN'), ...]

blob.noun_phrases   # WordList(['titular threat', 'blob',
                    #            'ultimate movie monster',
                    #            'amoeba-like mass', ...])

for sentence in blob.sentences:
    print(sentence.sentiment.polarity)
# 0.060
# -0.341
```

https://textblob.readthedocs.io

BERT Sequence-level tasks

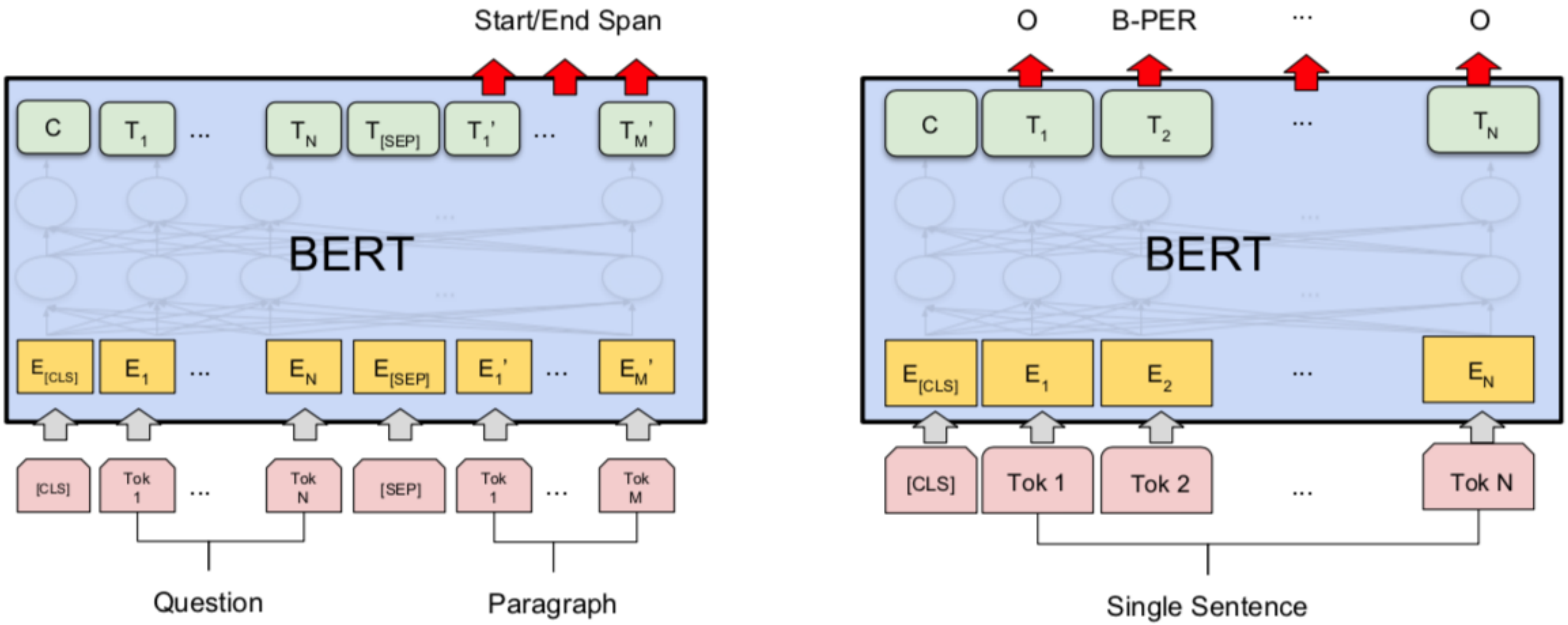


(a) Sentence Pair Classification Tasks:
MNLI, QQP, QNLI, STS-B, MRPC,
RTE, SWAG

(b) Single Sentence Classification Tasks:
SST-2, CoLA

Source: Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2018).
"BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." arXiv preprint arXiv:1810.04805

BERT Token-level tasks

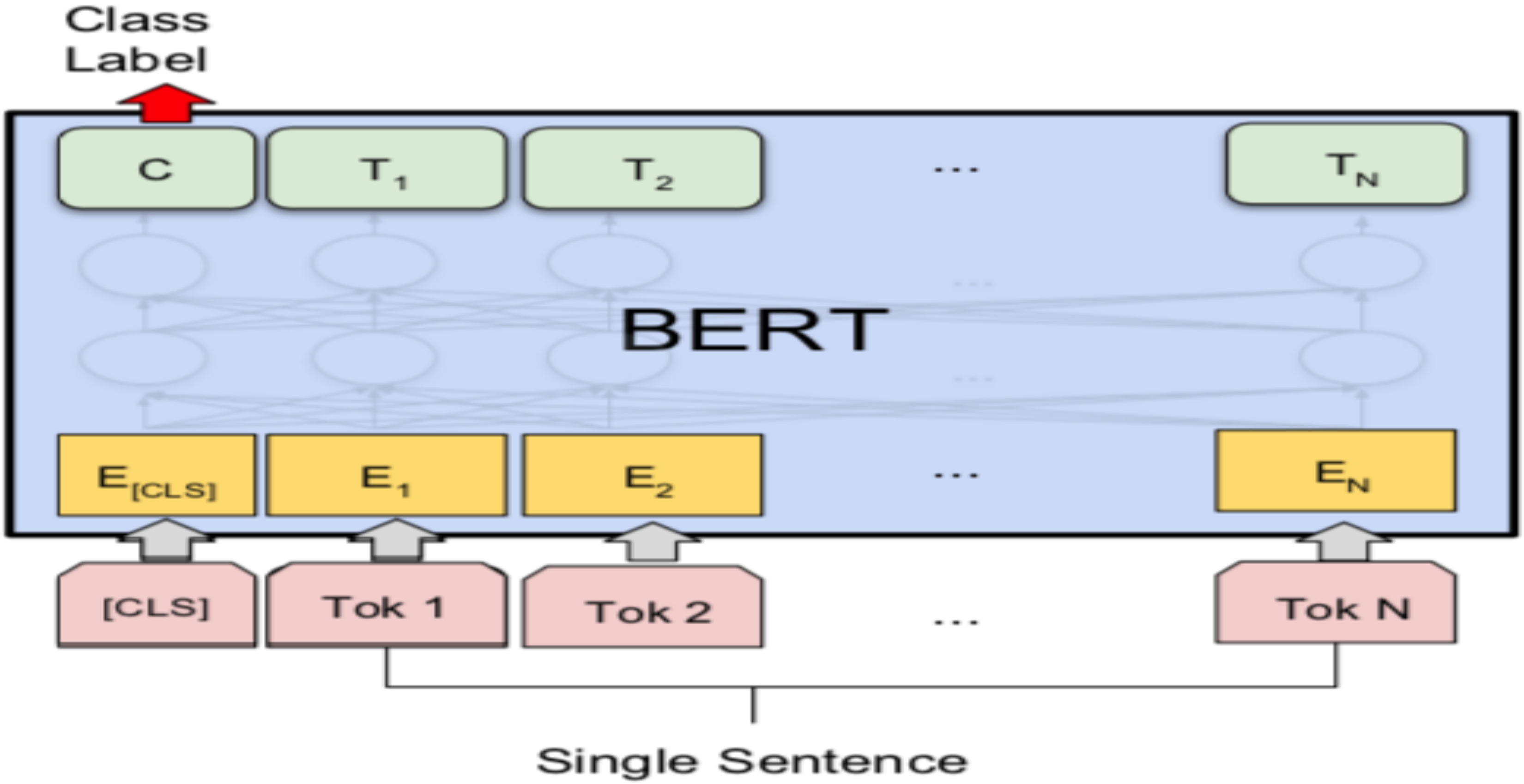


(c) Question Answering Tasks:
SQuAD v1.1

(d) Single Sentence Tagging Tasks:
CoNLL-2003 NER

Source: Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2018).
"BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." arXiv preprint arXiv:1810.04805

Sentiment Analysis: Single Sentence Classification

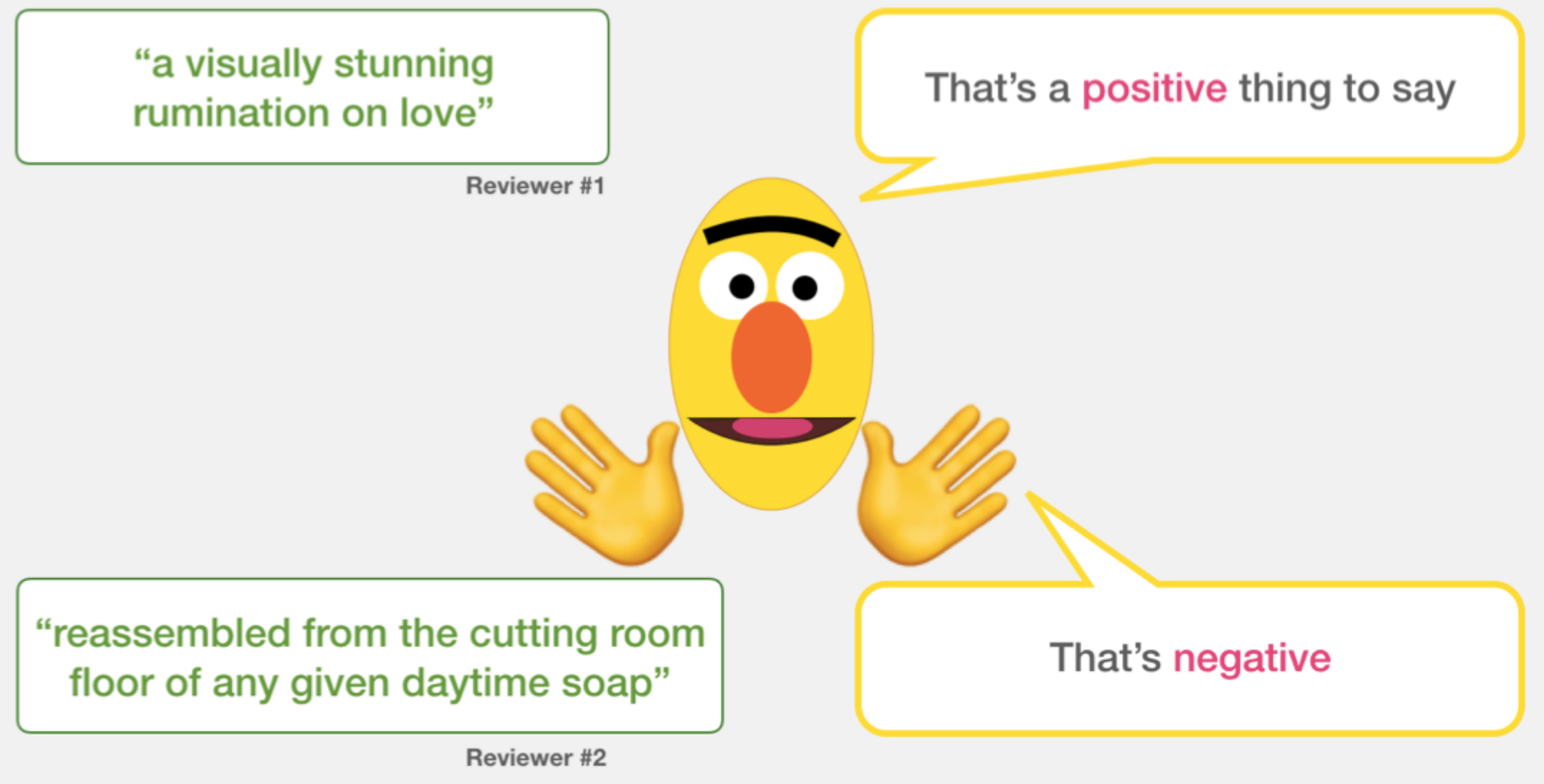


(b) Single Sentence Classification Tasks: SST-2, CoLA

Source: Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2018). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." arXiv preprint arXiv:1810.04805

A Visual Guide to Using BERT for the First Time

(Jay Alammar, 2019)


"a visually stunning rumination on love"
Reviewer #1
That's a positive thing to say
"reassembled from the cutting room floor of any given daytime soap"
Reviewer #2
That's negative

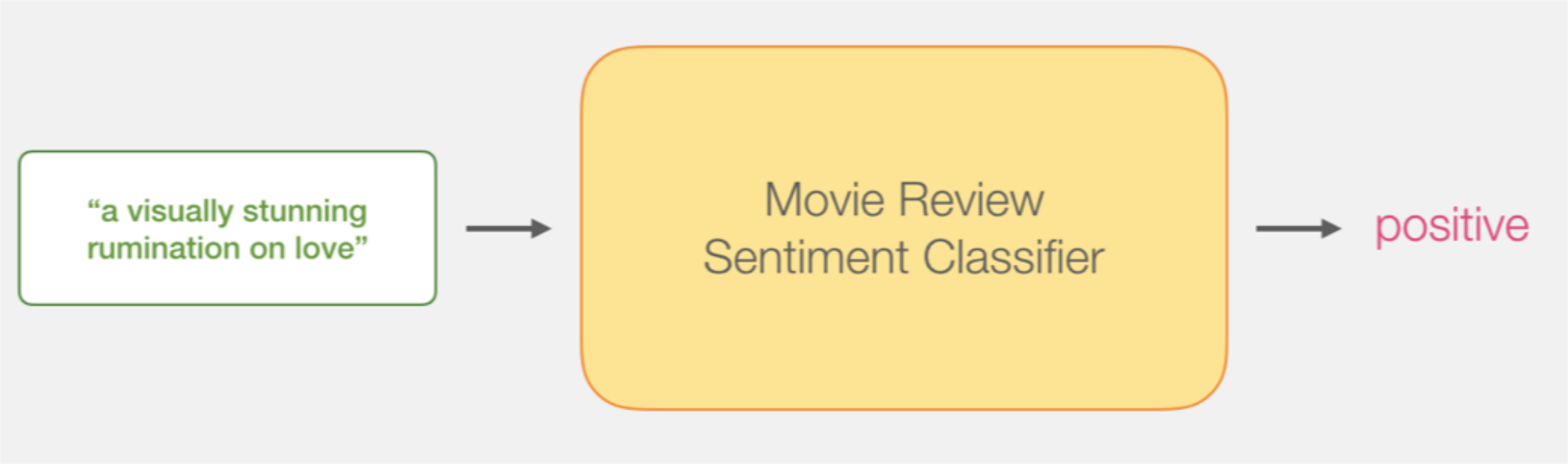
Source: Jay Alammar (2019), A Visual Guide to Using BERT for the First Time, http://jalammar.github.io/a-visual-guide-to-using-bert-for-the-first-time/

Sentiment Classification: SST2
Sentences from movie reviews

sentence	label
a stirring , funny and finally transporting re imagining of beauty and the beast and 1930s horror films	1
apparently reassembled from the cutting room floor of any given daytime soap	0
they presume their audience won't sit still for a sociology lesson	0
this is a visually stunning rumination on love , memory , history and the war between art and commerce	1
jonathan parker 's bartleby should have been the be all end all of the modern office anomie films	1

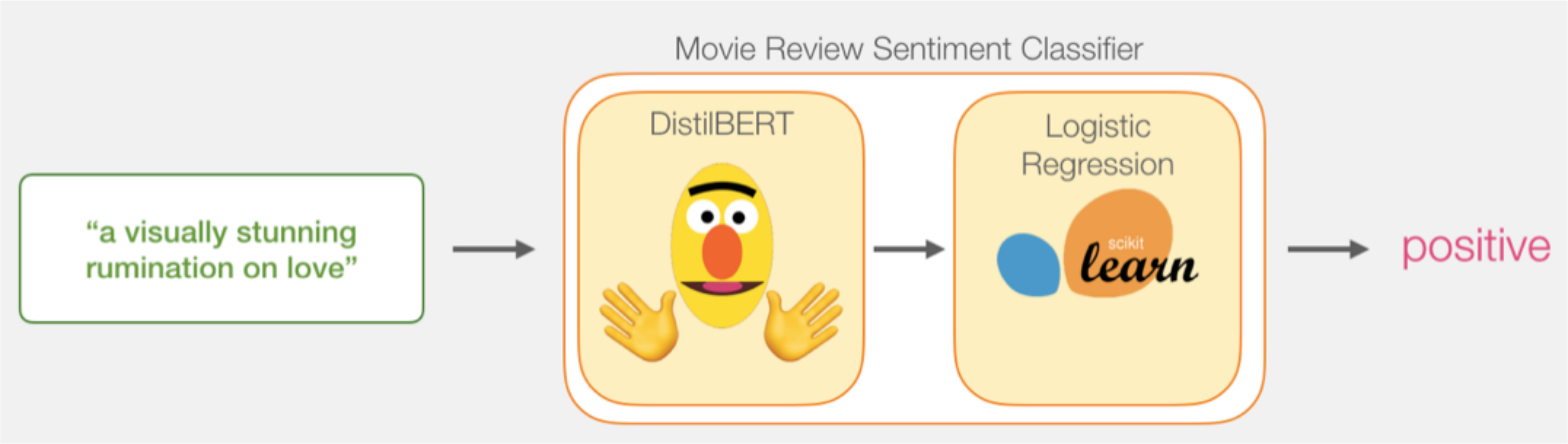
Source: Jay Alammar (2019), A Visual Guide to Using BERT for the First Time,
http://jalammar.github.io/a-visual-guide-to-using-bert-for-the-first-time/

Movie Review Sentiment Classifier


"a visually stunning rumination on love"
Movie Review Sentiment Classifier
positive

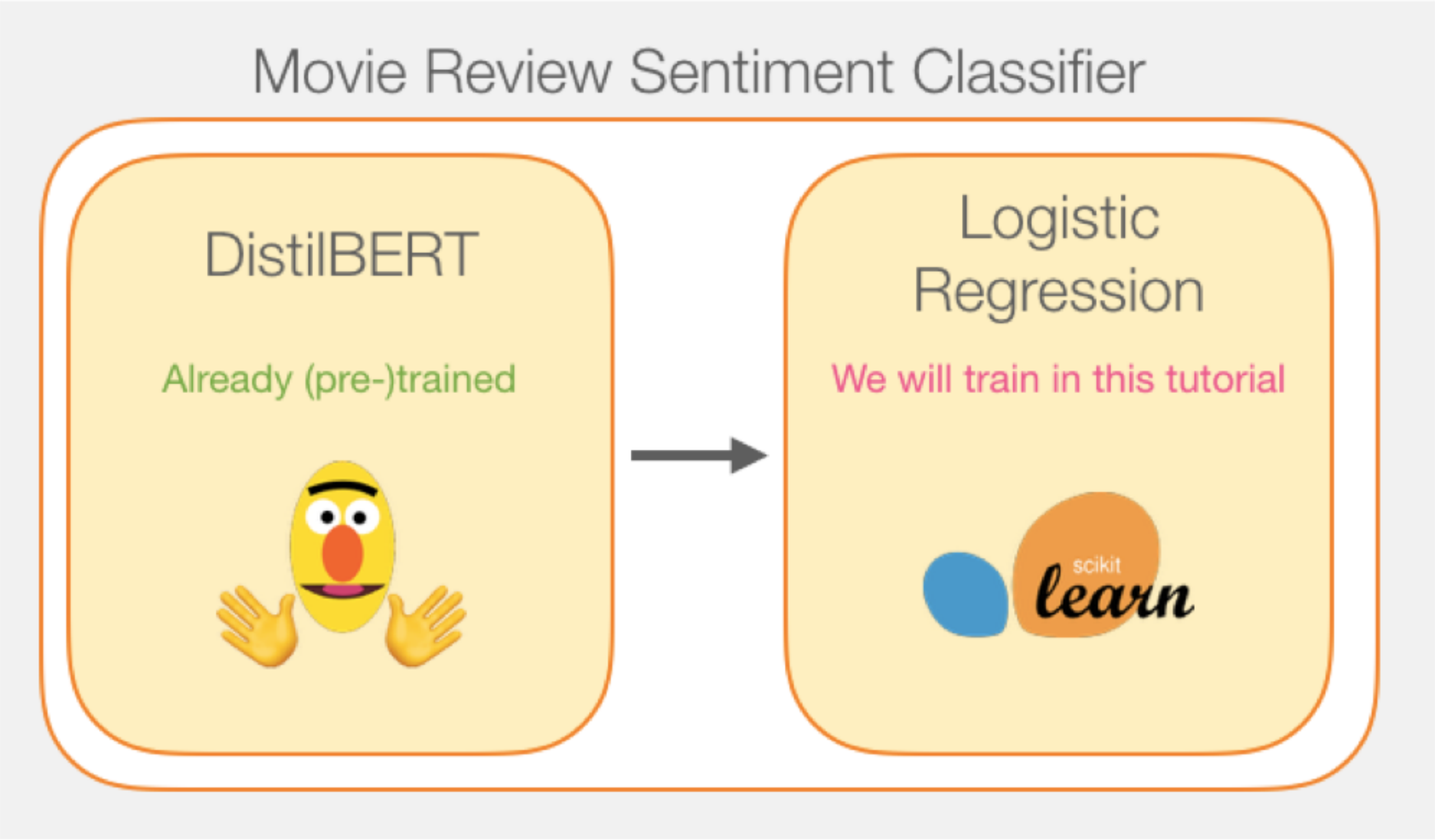
Source: Jay Alammar (2019), A Visual Guide to Using BERT for the First Time, http://jalammar.github.io/a-visual-guide-to-using-bert-for-the-first-time/

Movie Review Sentiment Classifier


Movie Review Sentiment Classifier
"a visually stunning rumination on love"
DistilBERT
Logistic Regression
scikit learn
positive

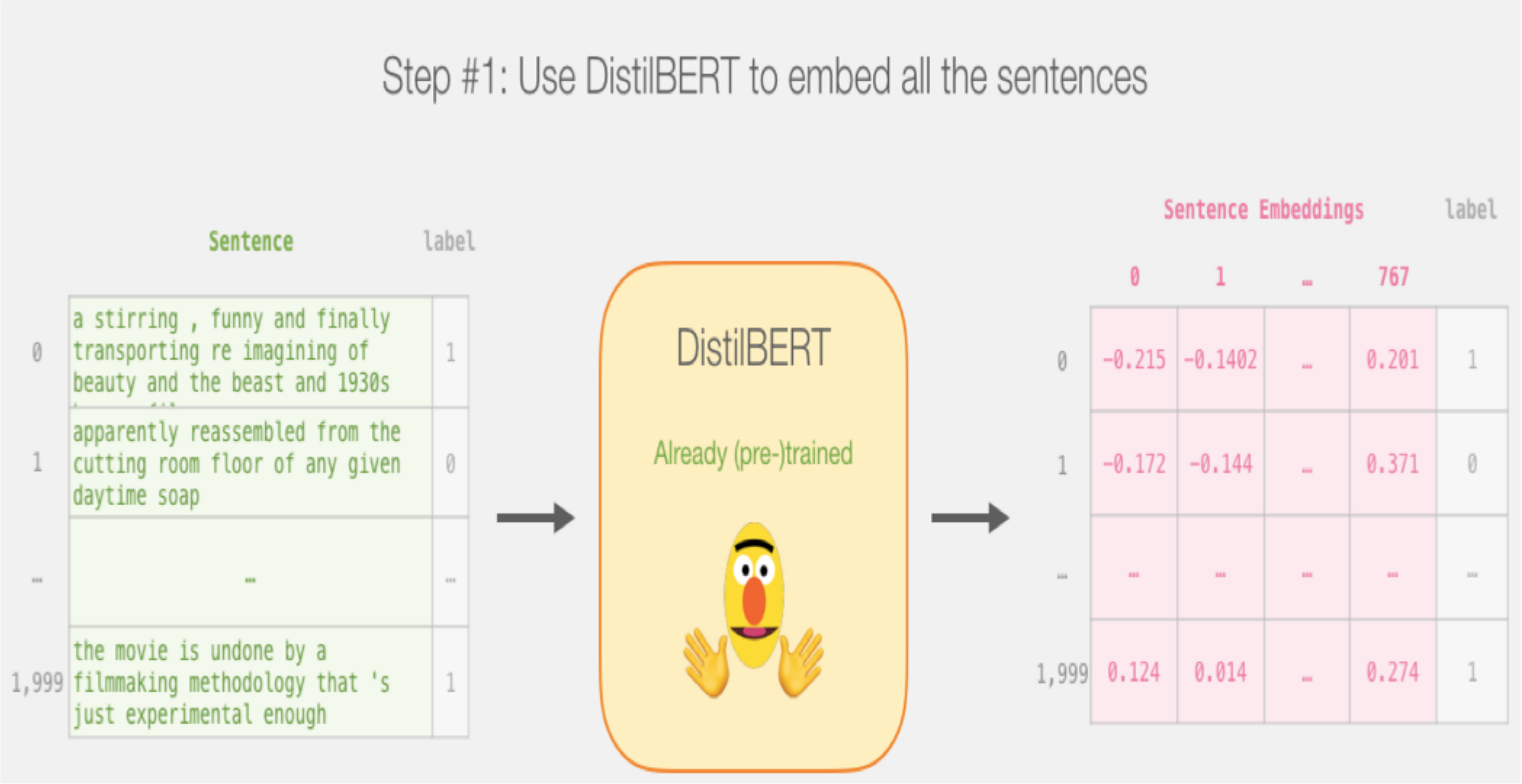
Source: Jay Alammar (2019), A Visual Guide to Using BERT for the First Time, http://jalammar.github.io/a-visual-guide-to-using-bert-for-the-first-time/

Movie Review Sentiment Classifier Model Training


Movie Review Sentiment Classifier
DistilBERT
Already (pre-)trained
Logistic Regression
We will train in this tutorial
scikit learn

Source: Jay Alammar (2019), A Visual Guide to Using BERT for the First Time, http://jalammar.github.io/a-visual-guide-to-using-bert-for-the-first-time/

Step # 1 Use distilBERT to Generate Sentence Embeddings



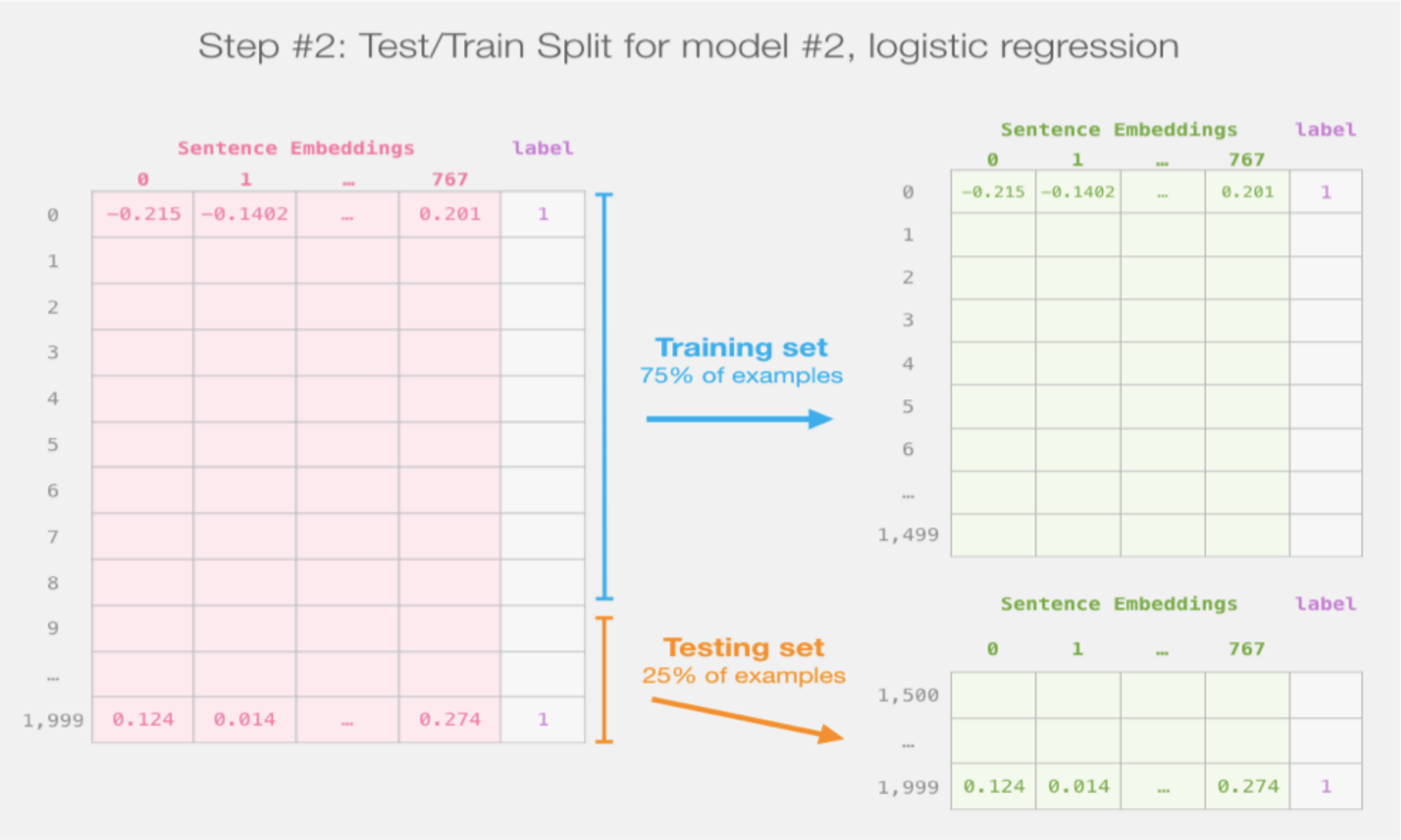
Step #1: Use DistilBERT to embed all the sentences

	Sentence	label
0	a stirring , funny and finally transporting re imagining of beauty and the beast and 1930s	1
1	apparently reassembled from the cutting room floor of any given daytime soap	0
…	…	…
1,999	the movie is undone by a filmmaking methodology that 's just experimental enough	1

	Sentence Embeddings				label
	0	1	…	767	
0	-0.215	-0.1402	…	0.201	1
1	-0.172	-0.144	…	0.371	0
…	…	…	…	…	…
1,999	0.124	0.014	…	0.274	1

Source: Jay Alammar (2019), A Visual Guide to Using BERT for the First Time, http://jalammar.github.io/a-visual-guide-to-using-bert-for-the-first-time/

Step #2:Test/Train Split for Model #2, Logistic Regression



Step #2: Test/Train Split for model #2, logistic regression

	Sentence Embeddings				label
	0	1	…	767	
0	-0.215	-0.1402	…	0.201	1
1					
2					
3					
4					
5					
6					
7					
8					
9					
…					
1,999	0.124	0.014	…	0.274	1

Training set
75% of examples

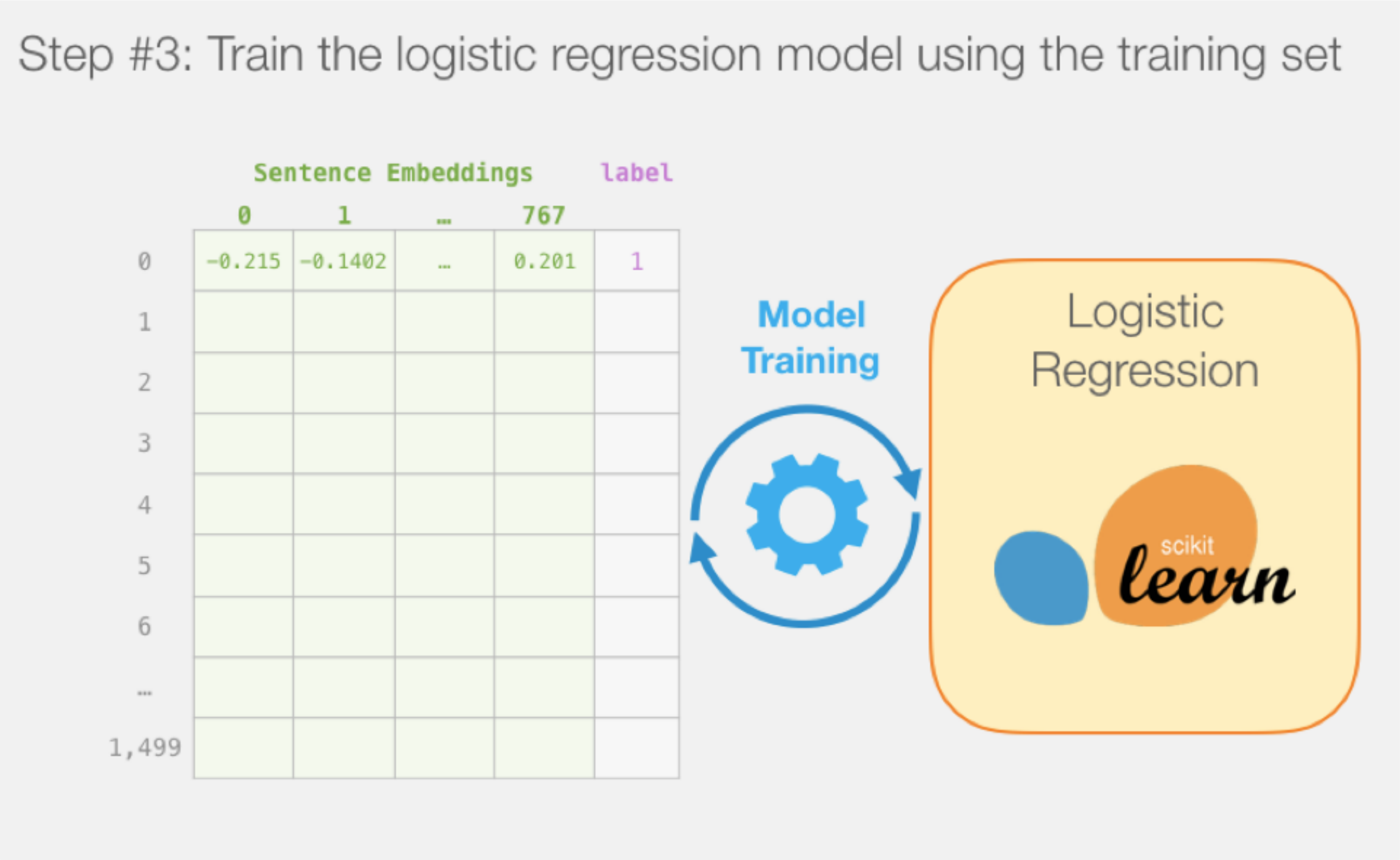
	Sentence Embeddings				label
	0	1	…	767	
0	-0.215	-0.1402	…	0.201	1
1					
2					
3					
4					
5					
6					
…					
1,499					

Testing set
25% of examples

	Sentence Embeddings				label
	0	1	…	767	
1,500					
…					
1,999	0.124	0.014	…	0.274	1

Source: Jay Alammar (2019), A Visual Guide to Using BERT for the First Time,
http://jalammar.github.io/a-visual-guide-to-using-bert-for-the-first-time/

Step #3 Train the logistic regression model using the training set



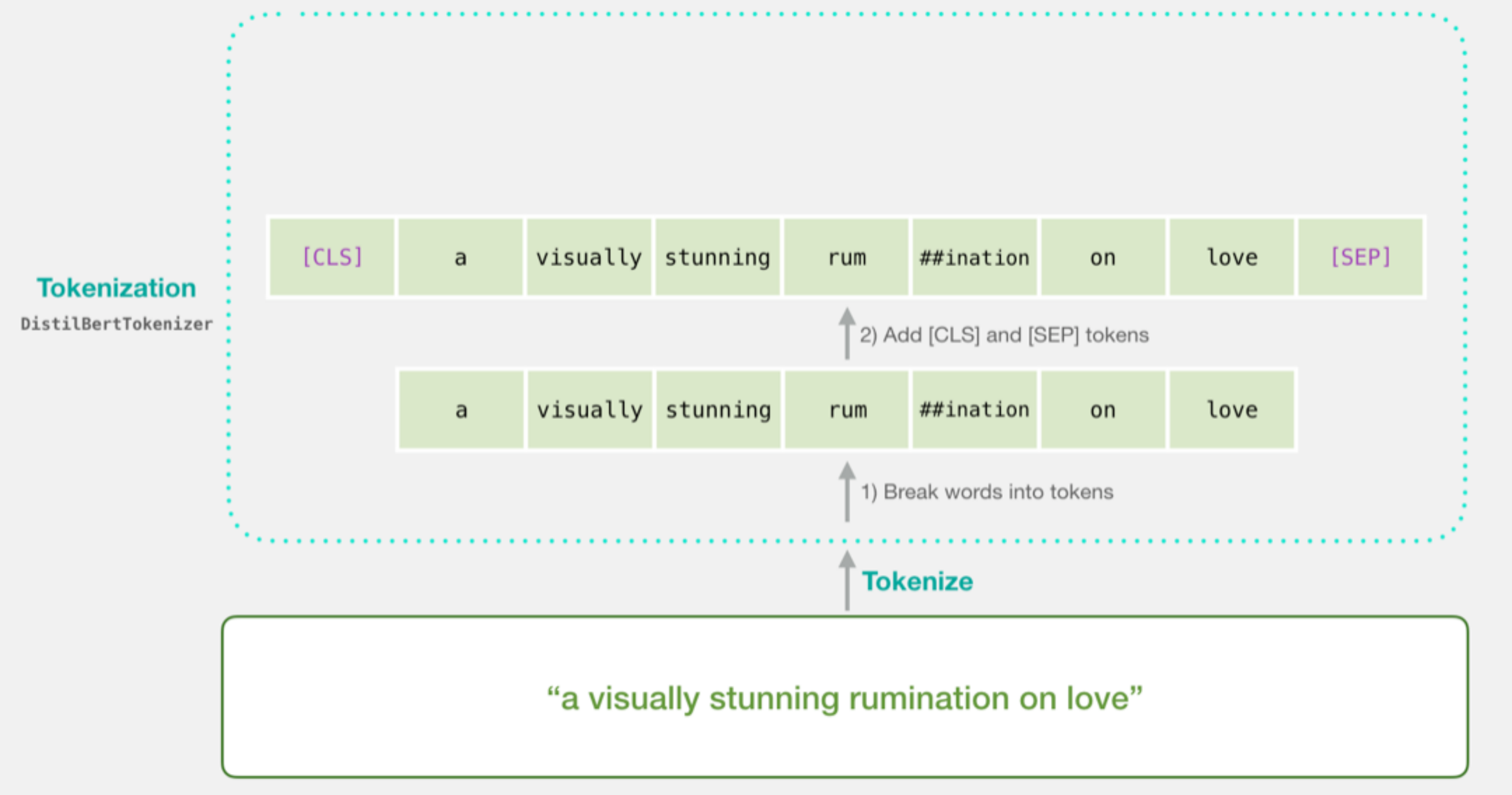
Step #3: Train the logistic regression model using the training set

	Sentence Embeddings				label
	0	1	…	767	
0	-0.215	-0.1402	…	0.201	1
1					
2					
3					
4					
5					
6					
…					
1,499					

Source: Jay Alammar (2019), A Visual Guide to Using BERT for the First Time, http://jalammar.github.io/a-visual-guide-to-using-bert-for-the-first-time/

Tokenization

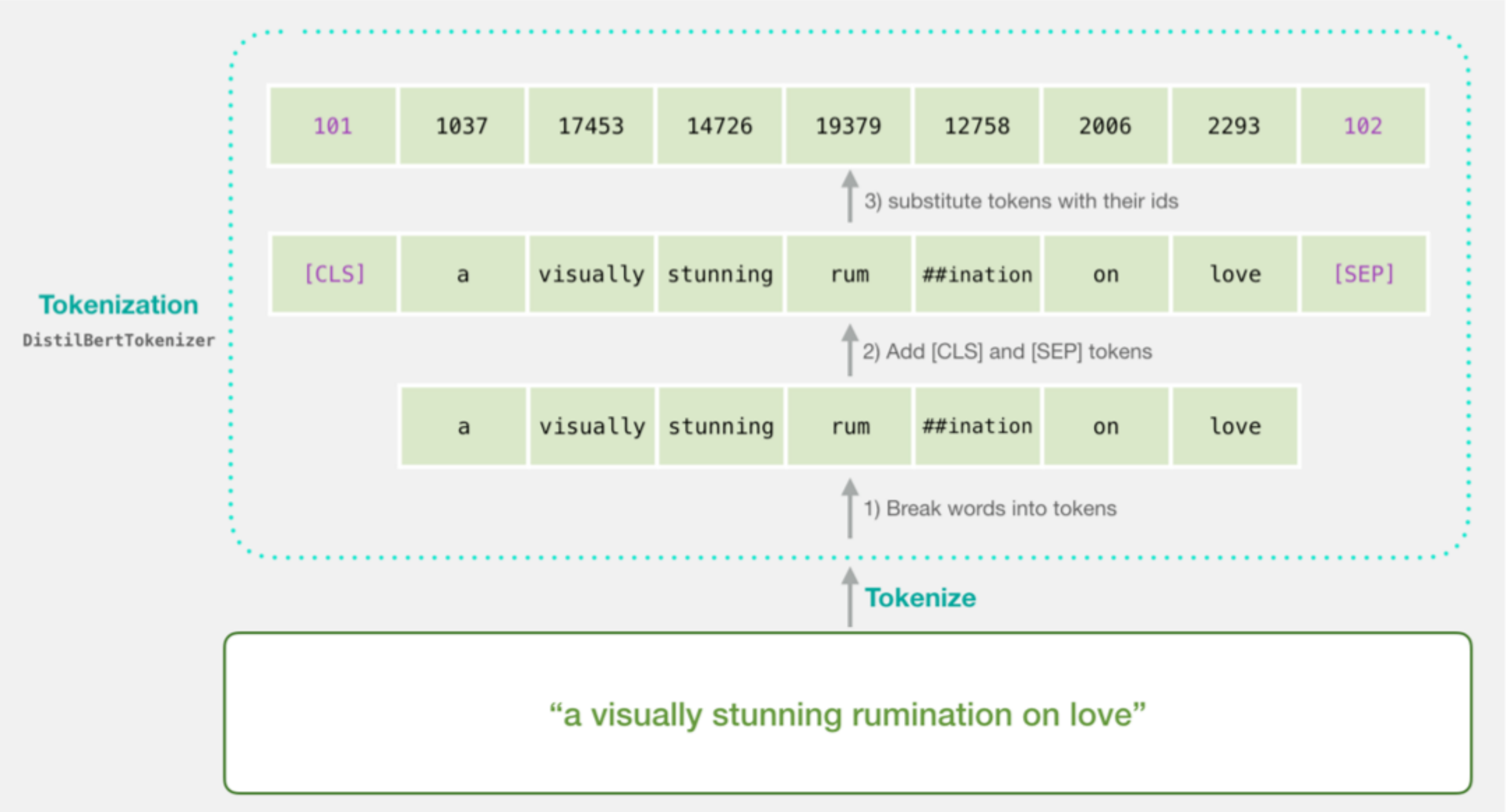
[CLS] a visually stunning rum ##ination on love [SEP]
a visually stunning rumination on love


Tokenization
DistilBertTokenizer
[CLS] a visually stunning rum ##ination on love [SEP]
2) Add [CLS] and [SEP] tokens
a visually stunning rum ##ination on love
1) Break words into tokens
Tokenize
"a visually stunning rumination on love"

Source: Jay Alammar (2019), A Visual Guide to Using BERT for the First Time,
http://jalammar.github.io/a-visual-guide-to-using-bert-for-the-first-time/

Tokenization

```
tokenizer.encode("a visually stunning rumination on love",
                 add_special_tokens=True)
```



Tokenization

DistilBertTokenizer

101	1037	17453	14726	19379	12758	2006	2293	102

3) substitute tokens with their ids

[CLS]	a	visually	stunning	rum	##ination	on	love	[SEP]

2) Add [CLS] and [SEP] tokens

a	visually	stunning	rum	##ination	on	love

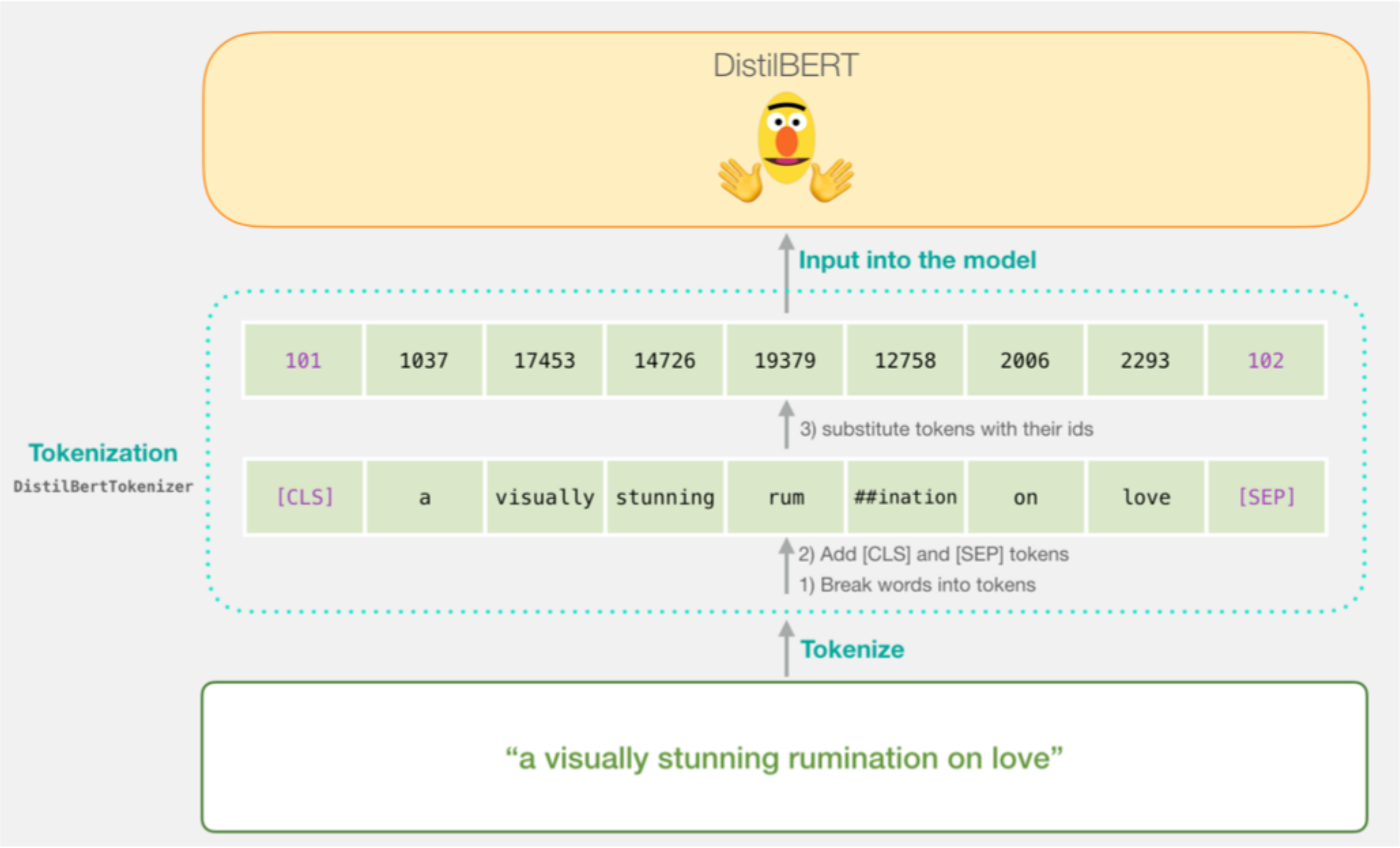
1) Break words into tokens

Tokenize

“a visually stunning rumination on love”

Source: Jay Alammar (2019), A Visual Guide to Using BERT for the First Time, http://jalammar.github.io/a-visual-guide-to-using-bert-for-the-first-time/

Tokenization for BERT Model



DistilBERT

Input into the model

101	1037	17453	14726	19379	12758	2006	2293	102
[CLS]	a	visually	stunning	rum	##ination	on	love	[SEP]

3) substitute tokens with their ids

2) Add [CLS] and [SEP] tokens

1) Break words into tokens

Tokenization

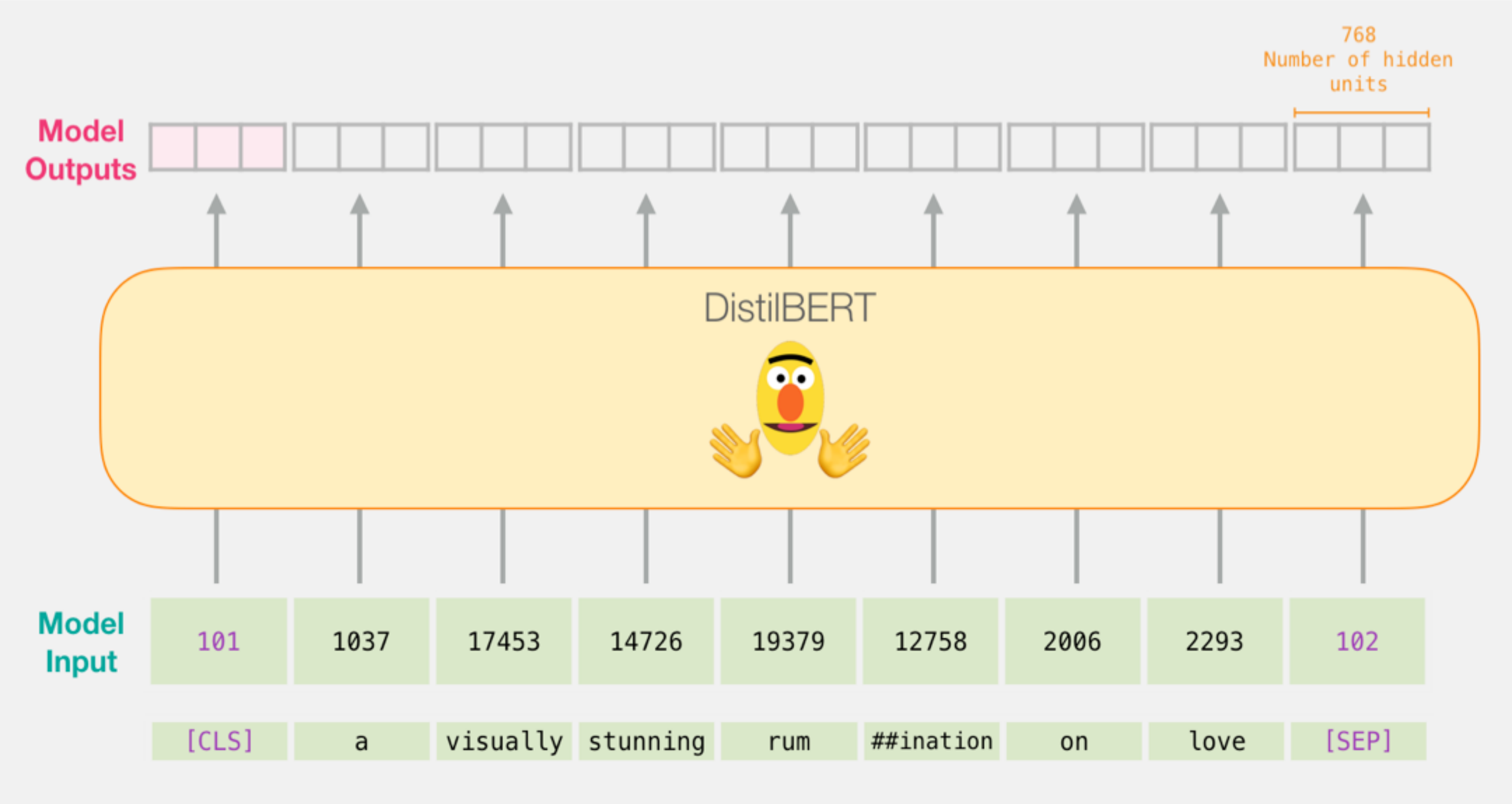
DistilBertTokenizer

Tokenize

“a visually stunning rumination on love”

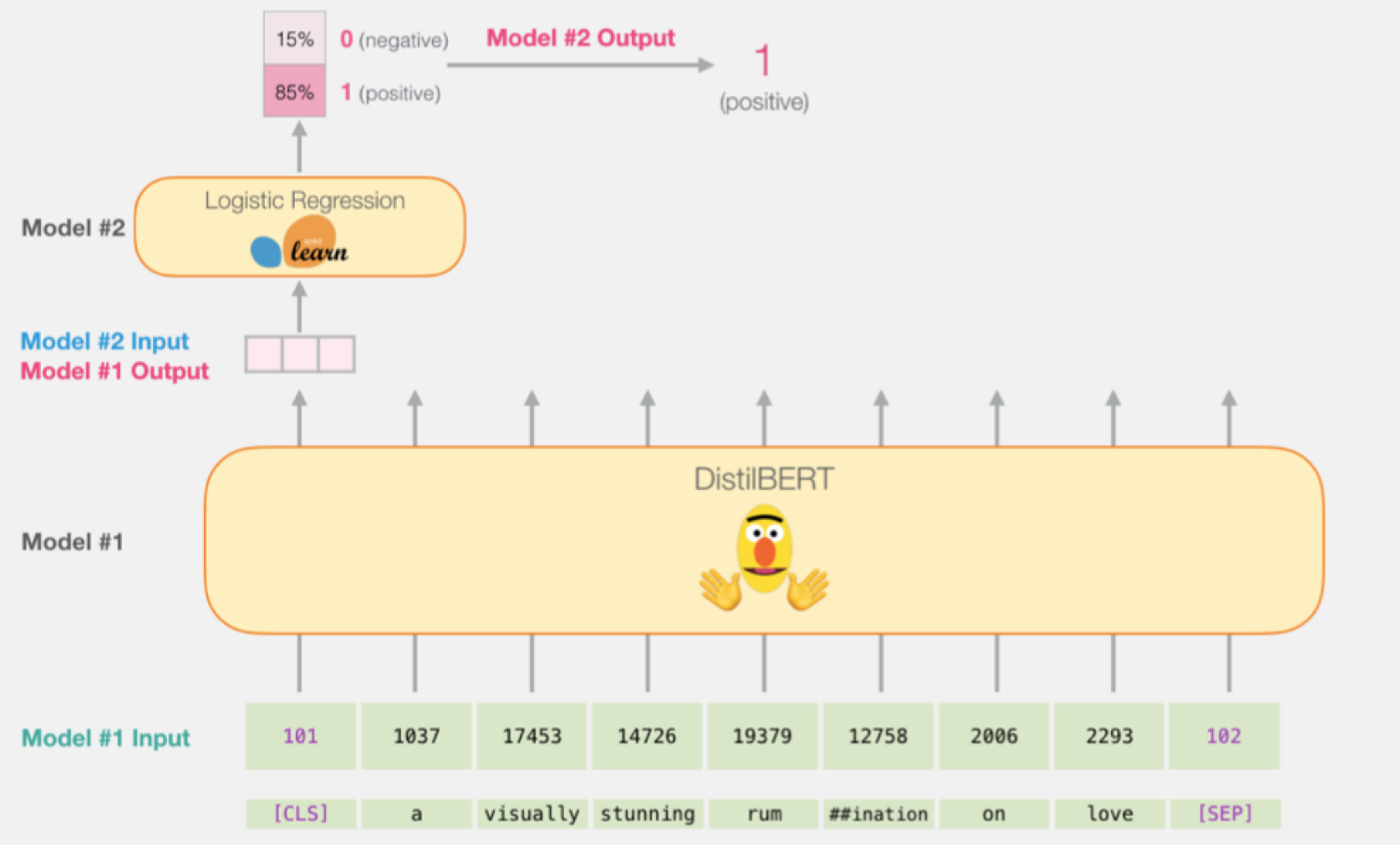
Source: Jay Alammar (2019), A Visual Guide to Using BERT for the First Time, http://jalammar.github.io/a-visual-guide-to-using-bert-for-the-first-time/

Flowing Through DistilBERT (768 features)



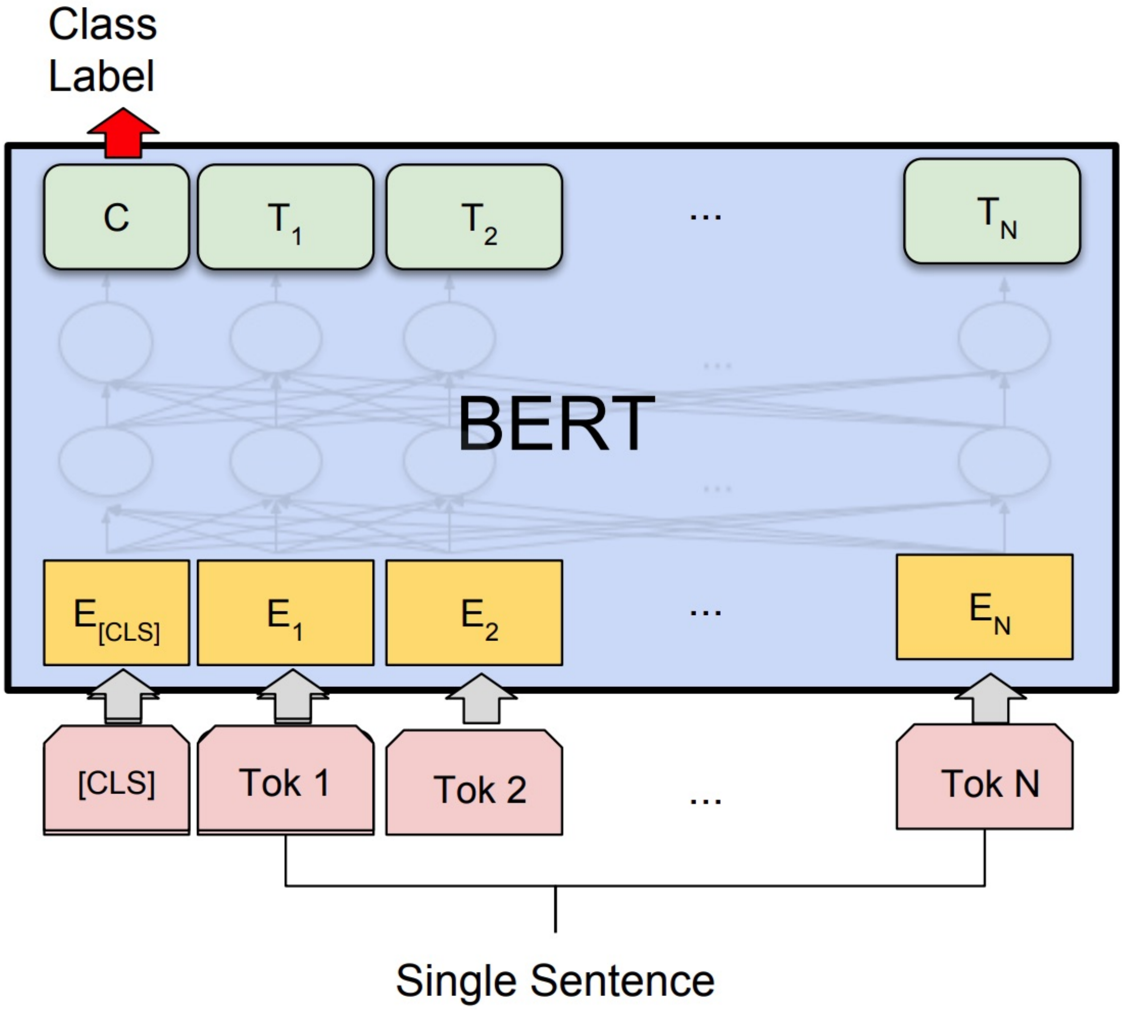
Source: Jay Alammar (2019), A Visual Guide to Using BERT for the First Time, http://jalammar.github.io/a-visual-guide-to-using-bert-for-the-first-time/

Model #1 Output Class vector as Model #2 Input



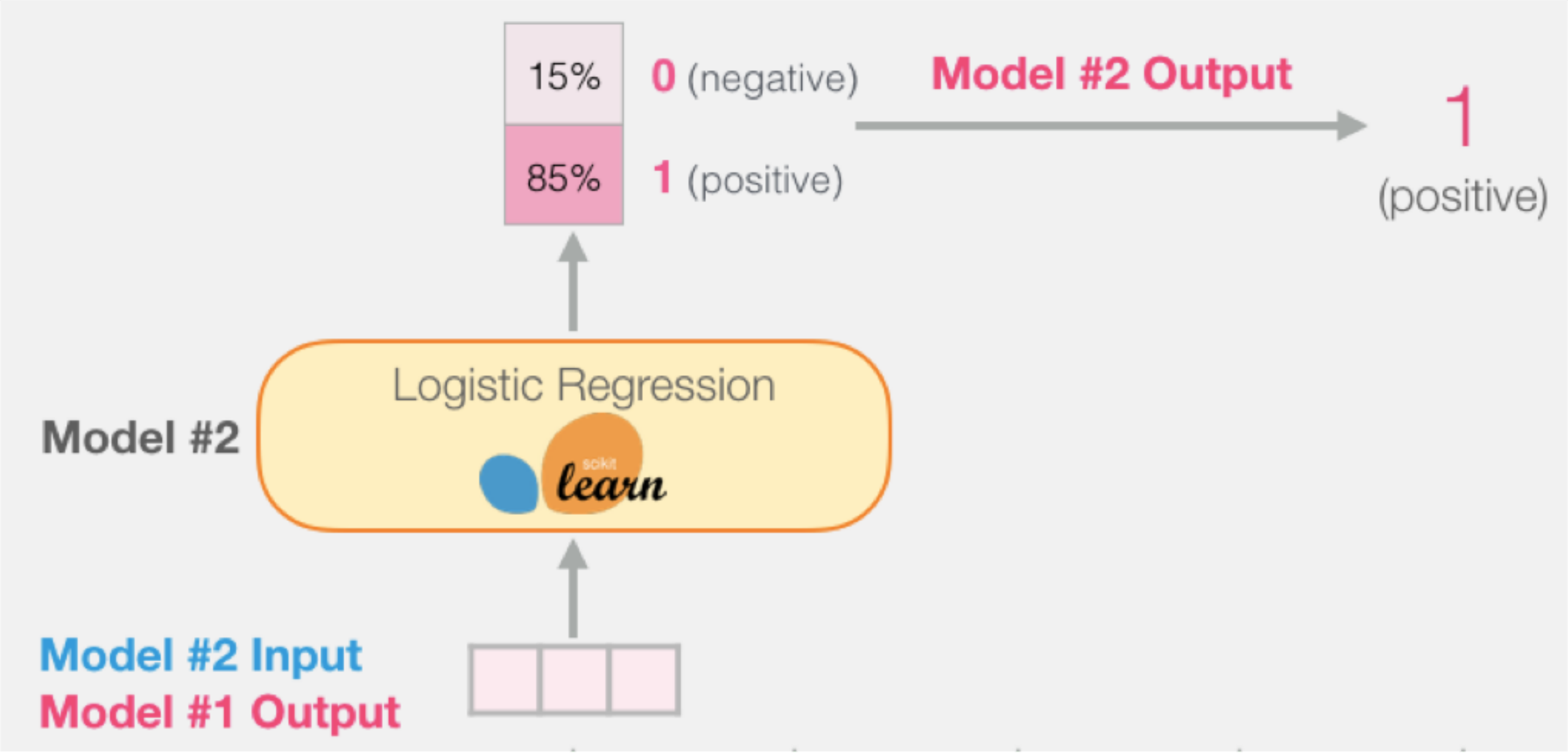
Source: Jay Alammar (2019), A Visual Guide to Using BERT for the First Time, http://jalammar.github.io/a-visual-guide-to-using-bert-for-the-first-time/

Fine-tuning BERT on Single Sentence Classification Tasks



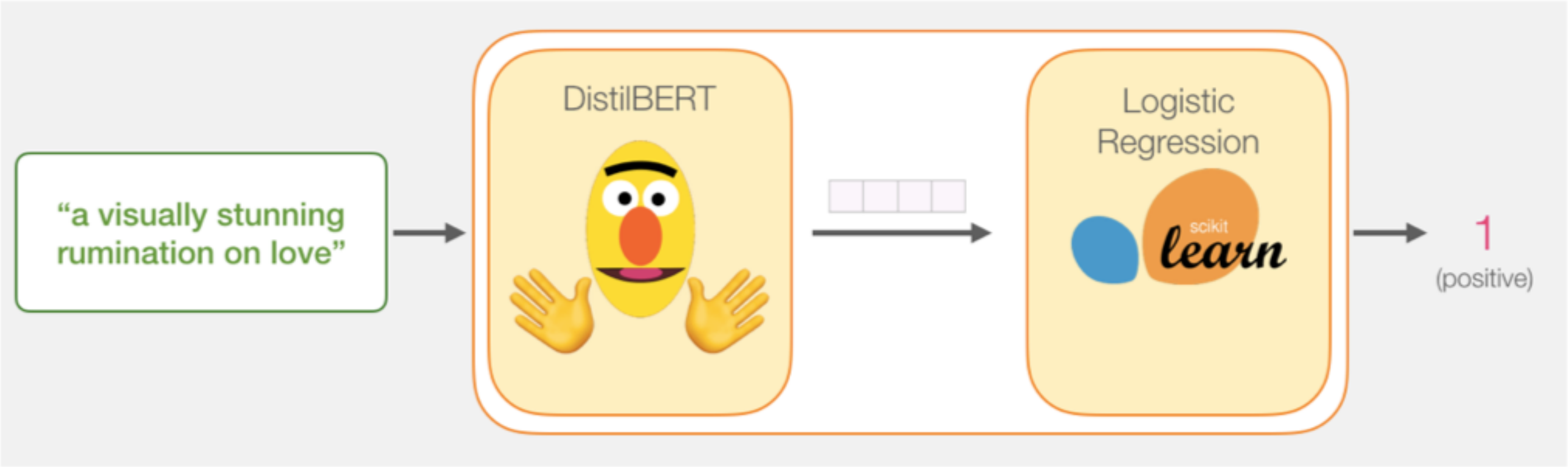
Source: Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2018). "Bert: Pre-training of deep bidirectional transformers for language understanding." arXiv preprint arXiv:1810.04805.

Model #1 Output Class vector as Model #2 Input



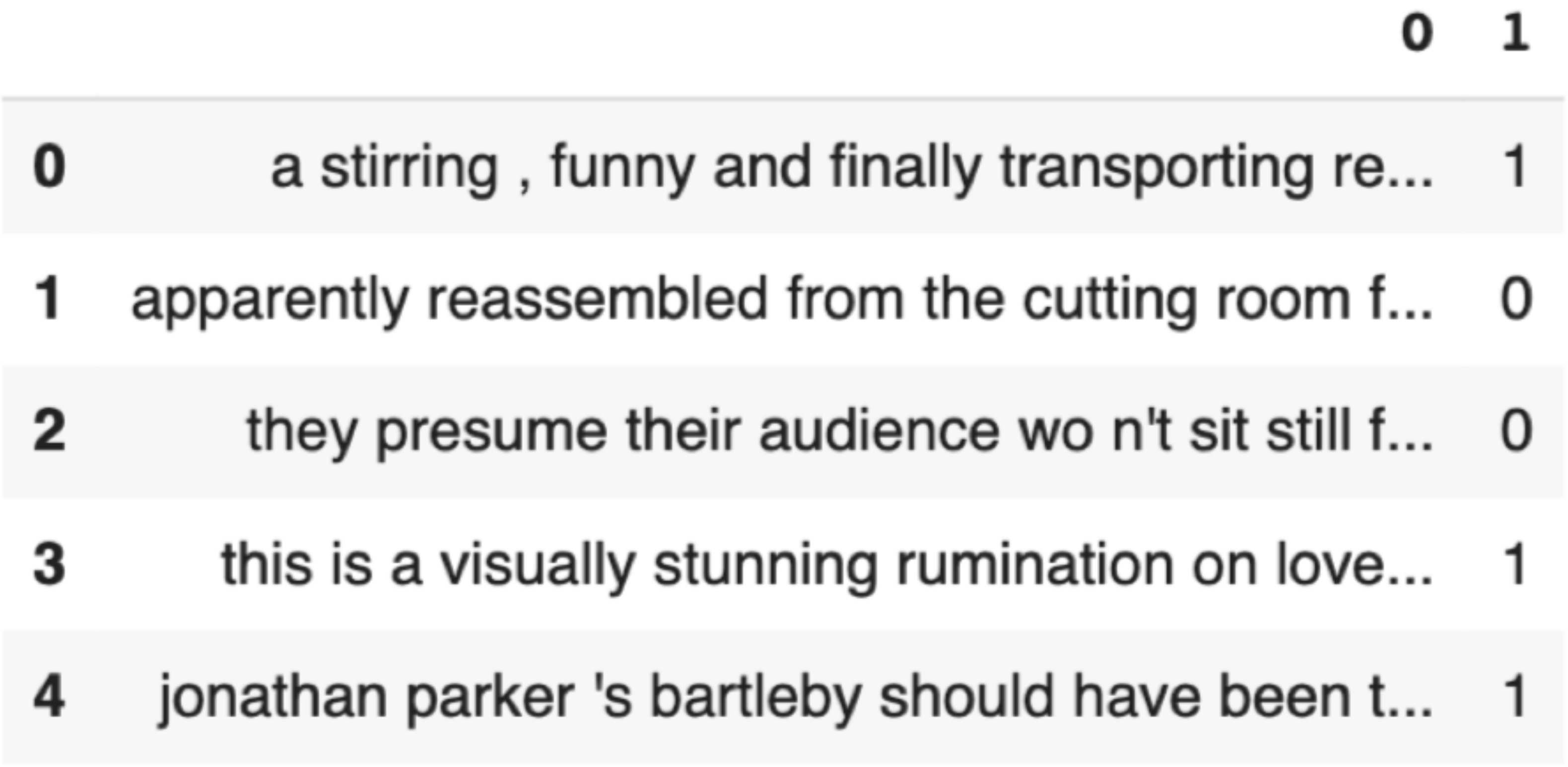
Source: Jay Alammar (2019), A Visual Guide to Using BERT for the First Time, http://jalammar.github.io/a-visual-guide-to-using-bert-for-the-first-time/

Logistic Regression Model to classify Class vector



Source: Jay Alammar (2019), A Visual Guide to Using BERT for the First Time, http://jalammar.github.io/a-visual-guide-to-using-bert-for-the-first-time/

```
df = pd.read_csv('https://github.com/clairett/pytorch-sentiment-classification/raw/master/data/SST2/train.tsv', delimiter='\t', header=None)

df.head()
```

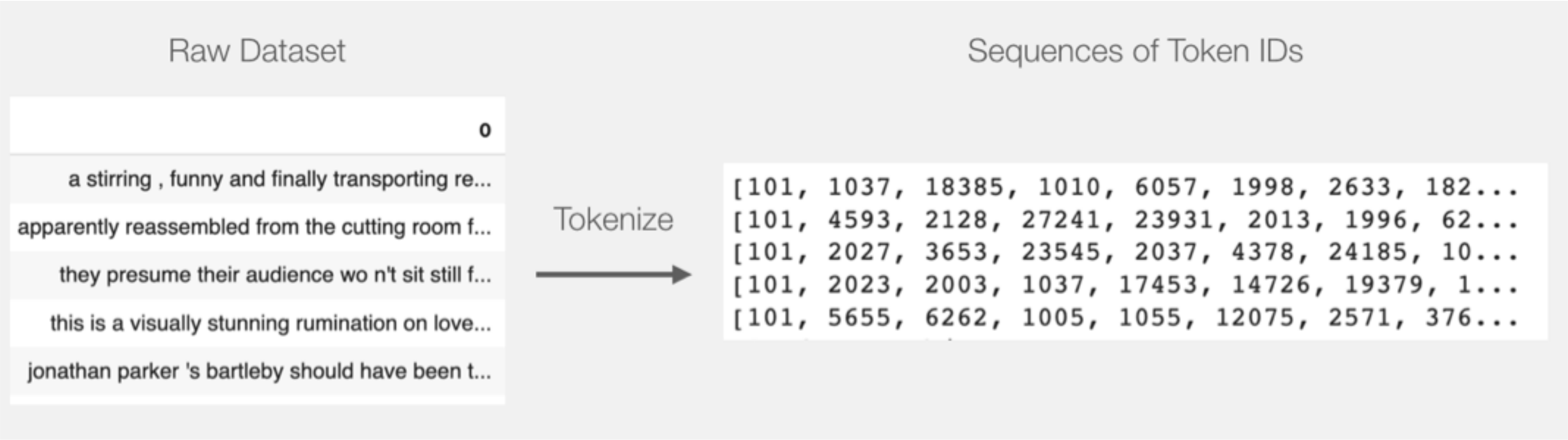


	0	1
0	a stirring , funny and finally transporting re...	1
1	apparently reassembled from the cutting room f...	0
2	they presume their audience wo n't sit still f...	0
3	this is a visually stunning rumination on love...	1
4	jonathan parker 's bartleby should have been t...	1

Source: Jay Alammar (2019), A Visual Guide to Using BERT for the First Time, http://jalammar.github.io/a-visual-guide-to-using-bert-for-the-first-time/

Tokenization

```
tokenized = df[0].apply((lambda x: tokenizer.encode(x,
add_special_tokens=True)))
```

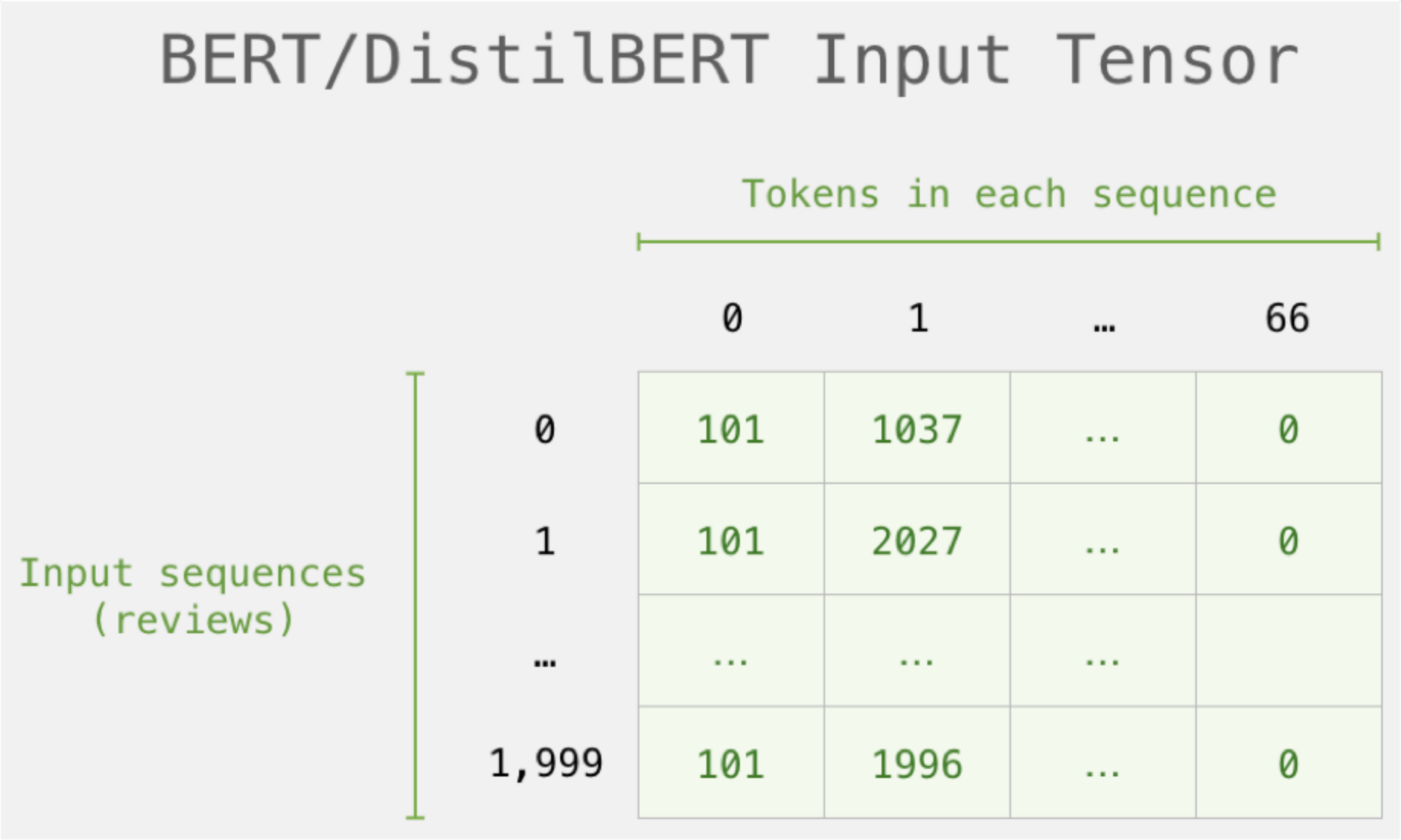


0
a stirring , funny and finally transporting re...
apparently reassembled from the cutting room f...
they presume their audience wo n't sit still f...
this is a visually stunning rumination on love...
jonathan parker 's bartleby should have been t...

```
[101, 1037, 18385, 1010, 6057, 1998, 2633, 182...
[101, 4593, 2128, 27241, 23931, 2013, 1996, 62...
[101, 2027, 3653, 23545, 2037, 4378, 24185, 10...
[101, 2023, 2003, 1037, 17453, 14726, 19379, 1...
[101, 5655, 6262, 1005, 1055, 12075, 2571, 376...
```

Source: Jay Alammar (2019), A Visual Guide to Using BERT for the First Time,
http://jalammar.github.io/a-visual-guide-to-using-bert-for-the-first-time/

BERT Input Tensor



BERT/DistilBERT Input Tensor

Tokens in each sequence

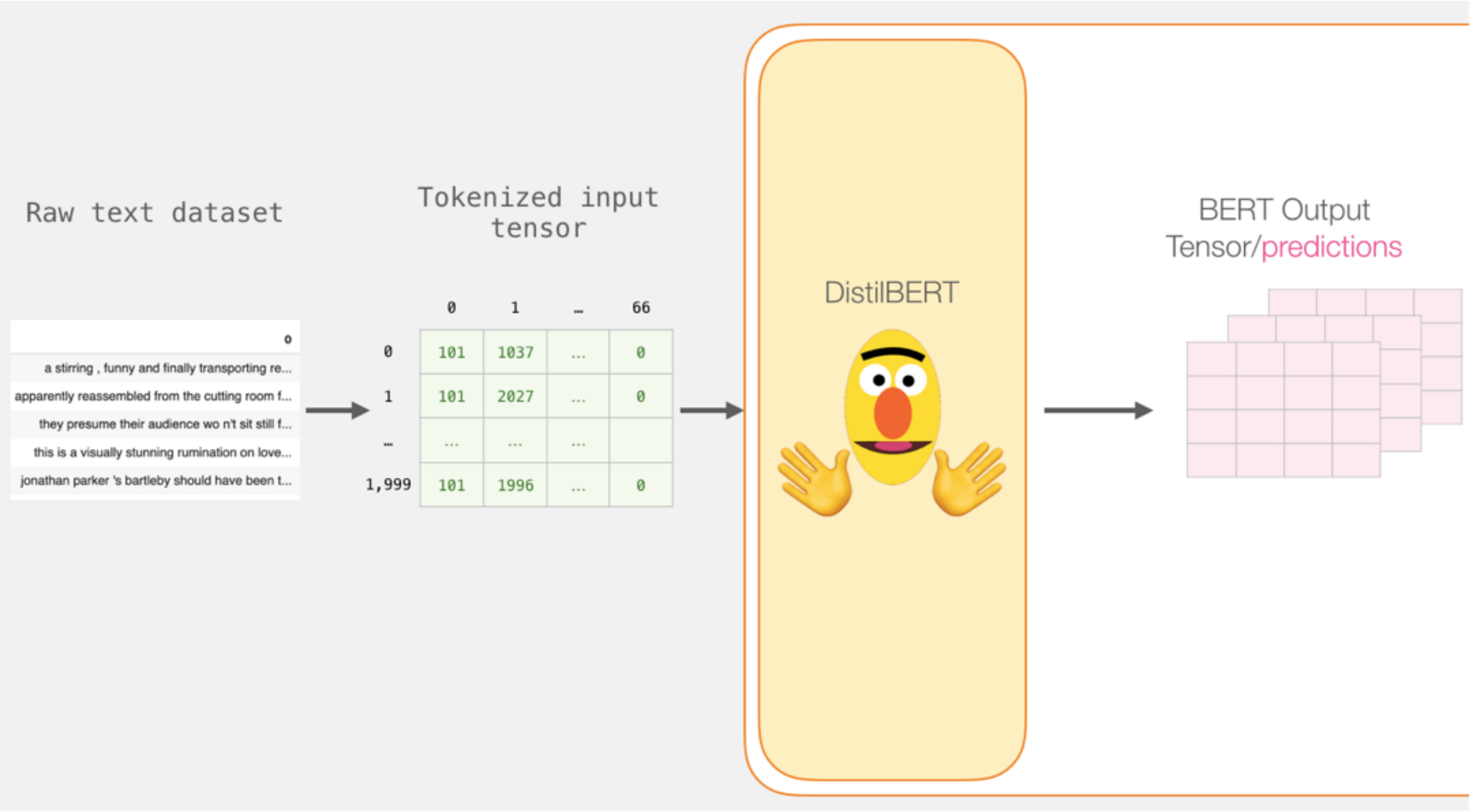
Input sequences (reviews)

	0	1	…	66
0	101	1037	...	0
1	101	2027	...	0
…	...	...	...	
1,999	101	1996	...	0

Source: Jay Alammar (2019), A Visual Guide to Using BERT for the First Time,
http://jalammar.github.io/a-visual-guide-to-using-bert-for-the-first-time/

Processing with DistilBERT

```
input_ids = torch.tensor(np.array(padded))
last_hidden_states = model(input_ids)
```

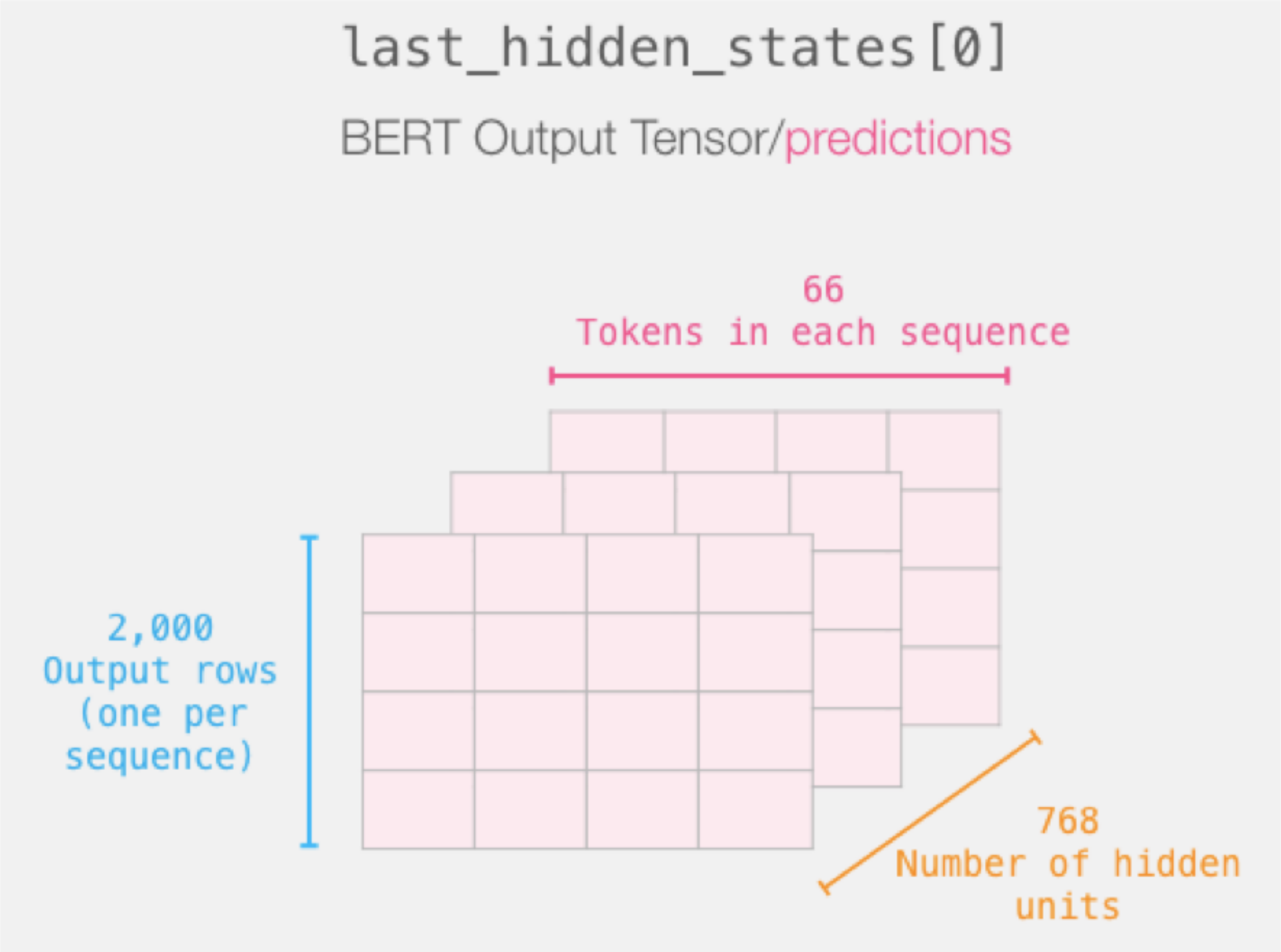


0
a stirring , funny and finally transporting re...
apparently reassembled from the cutting room f...
they presume their audience wo n't sit still f...
this is a visually stunning rumination on love...
jonathan parker 's bartleby should have been t...

	0	1	…	66
0	101	1037	...	0
1	101	2027	...	0
…	...	...	...	
1,999	101	1996	...	0

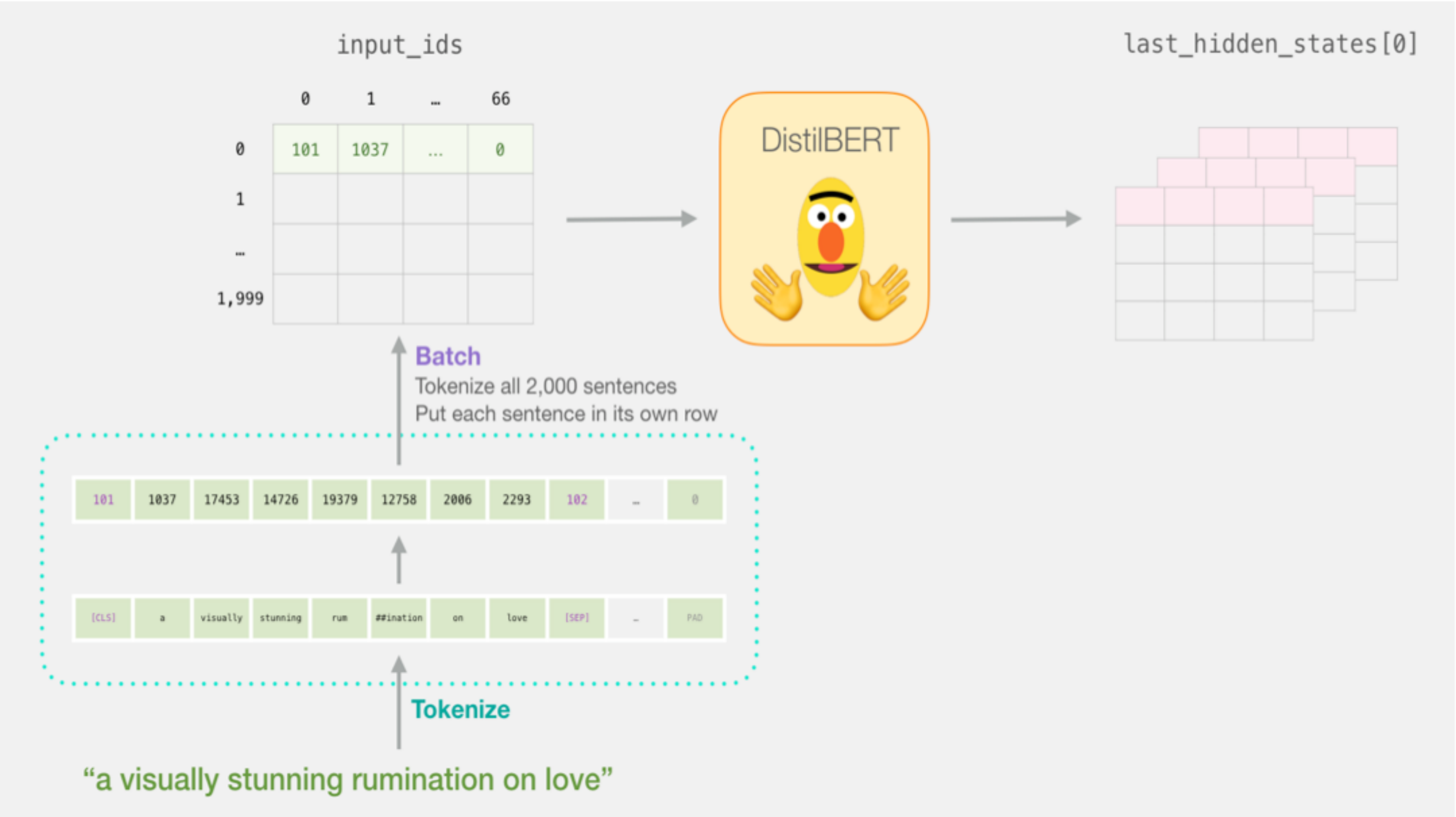
Source: Jay Alammar (2019), A Visual Guide to Using BERT for the First Time, http://jalammar.github.io/a-visual-guide-to-using-bert-for-the-first-time/

Unpacking the BERT output tensor



Source: Jay Alammar (2019), A Visual Guide to Using BERT for the First Time,
http://jalammar.github.io/a-visual-guide-to-using-bert-for-the-first-time/

Sentence to last_hidden_state[0]



	0	1	…	66
0	101	1037	…	0
1				
…				
1,999				

Batch
Tokenize all 2,000 sentences
Put each sentence in its own row

101	1037	17453	14726	19379	12758	2006	2293	102	…	0
[CLS]	a	visually	stunning	rum	##ination	on	love	[SEP]	…	PAD

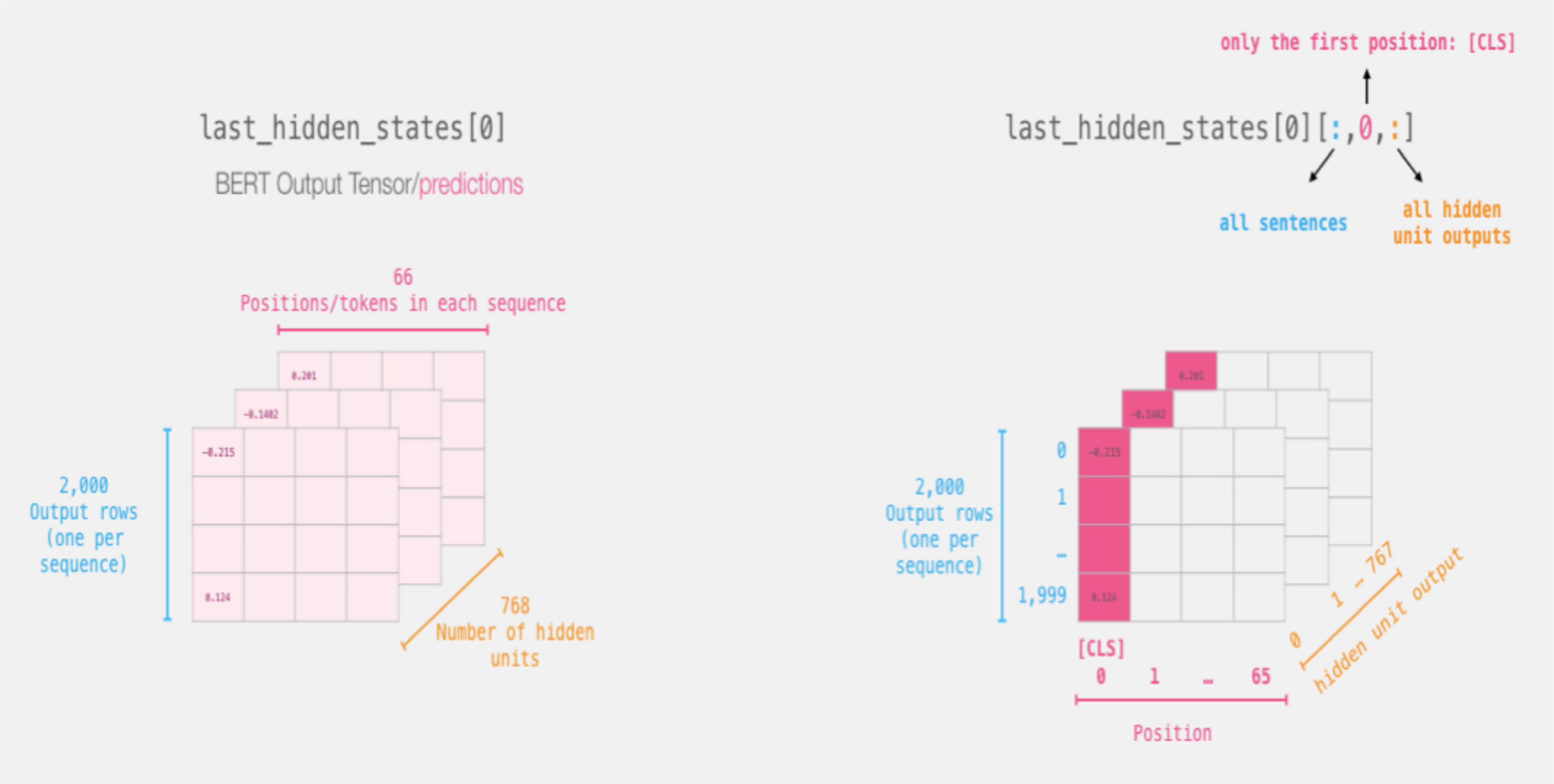
Tokenize

“a visually stunning rumination on love”

Source: Jay Alammar (2019), A Visual Guide to Using BERT for the First Time, http://jalammar.github.io/a-visual-guide-to-using-bert-for-the-first-time/

BERT’s output for the [CLS] tokens

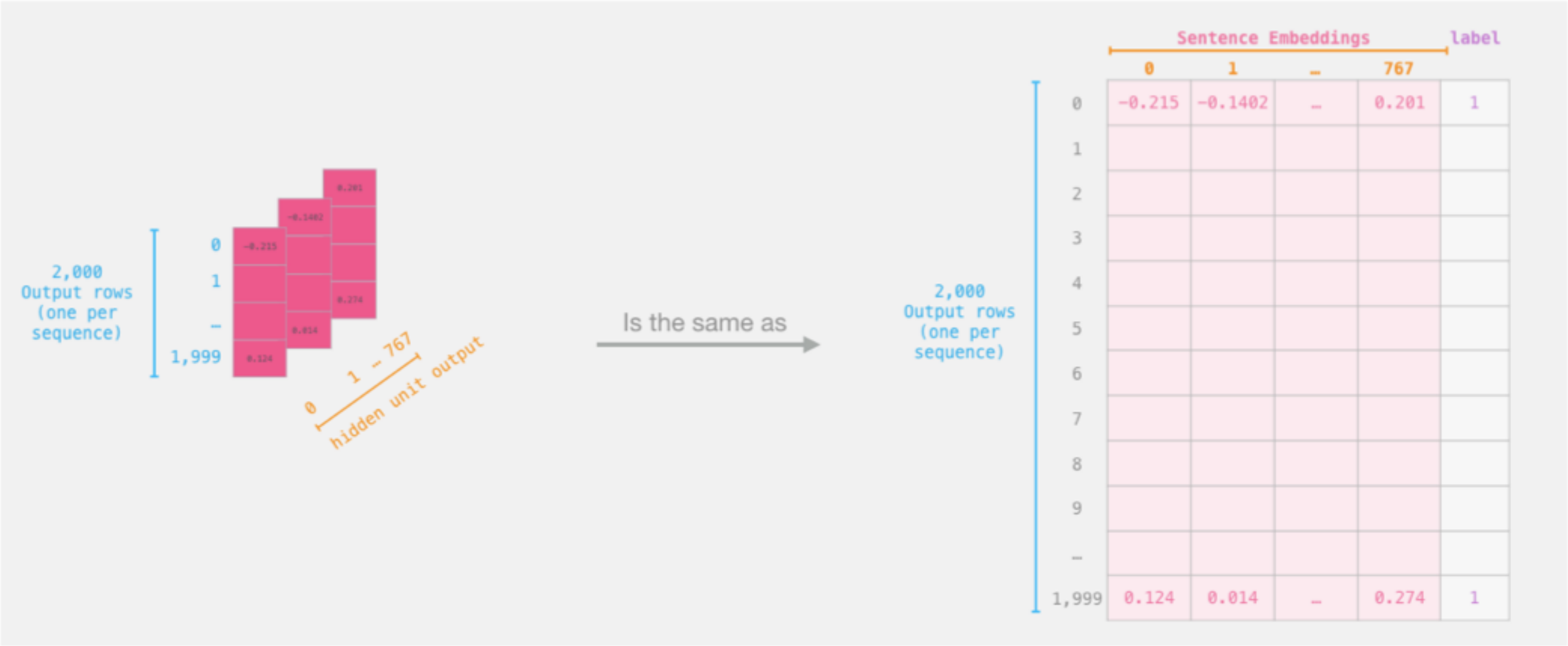
```
# Slice the output for the first position for all the
sequences, take all hidden unit outputs
features = last_hidden_states[0][:,0,:].numpy()
```



Source: Jay Alammar (2019), A Visual Guide to Using BERT for the First Time,
http://jalammar.github.io/a-visual-guide-to-using-bert-for-the-first-time/

The tensor sliced from BERT's output

Sentence Embeddings

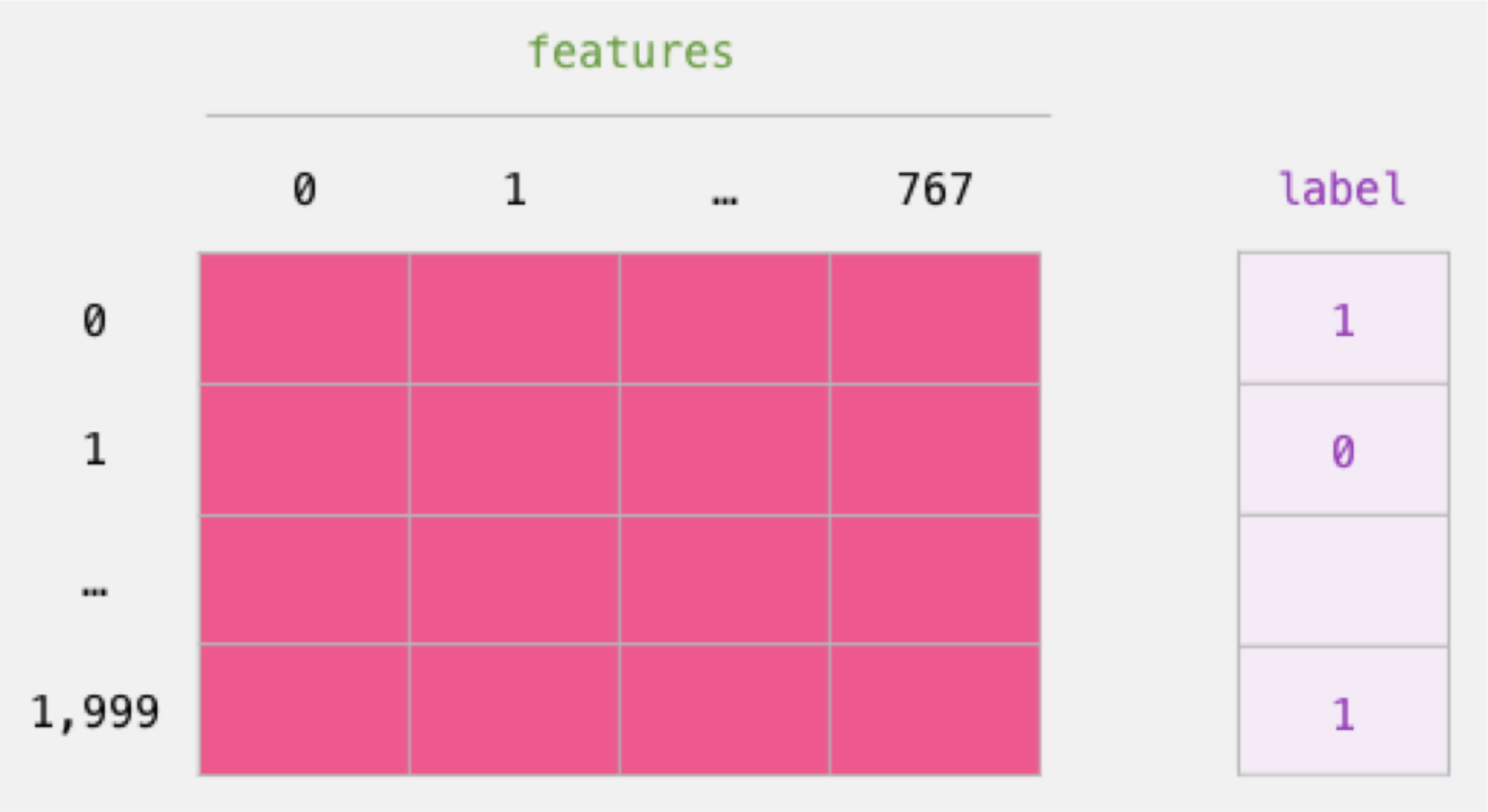


	Sentence Embeddings				label
	0	1	…	767	
0	-0.215	-0.1402	…	0.201	1
1					
2					
3					
4					
5					
6					
7					
8					
9					
…					
1,999	0.124	0.014	…	0.274	1

Source: Jay Alammar (2019), A Visual Guide to Using BERT for the First Time,
http://jalammar.github.io/a-visual-guide-to-using-bert-for-the-first-time/

Dataset for Logistic Regression (768 Features)

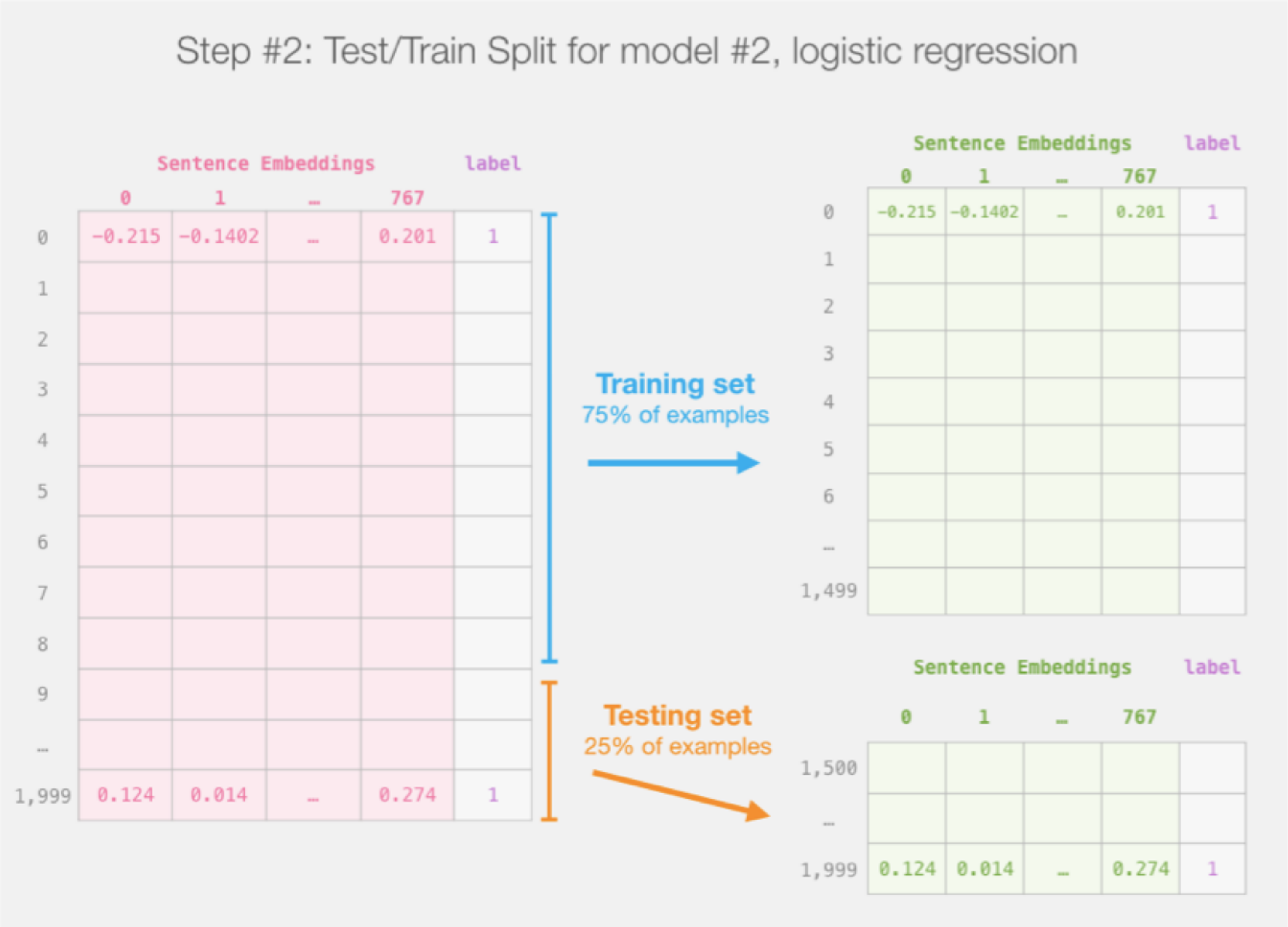
The features are the output vectors of BERT for the [CLS] token (position #0)



	features				
	0	1	…	767	label
0					1
1					0
…					
1,999					1

Source: Jay Alammar (2019), A Visual Guide to Using BERT for the First Time, http://jalammar.github.io/a-visual-guide-to-using-bert-for-the-first-time/

```
labels = df[1]
train_features, test_features, train_labels, test_labels = 
train_test_split(features, labels)
```



Step #2: Test/Train Split for model #2, logistic regression

	Sentence Embeddings				label
	0	1	…	767	
0	-0.215	-0.1402	…	0.201	1
1					
2					
3					
4					
5					
6					
7					
8					
9					
…					
1,999	0.124	0.014	…	0.274	1

Training set
75% of examples

	Sentence Embeddings				label
	0	1	…	767	
0	-0.215	-0.1402	…	0.201	1
1					
2					
3					
4					
5					
6					
…					
1,499					

Testing set
25% of examples

	Sentence Embeddings				label
	0	1	…	767	
1,500					
…					
1,999	0.124	0.014	…	0.274	1

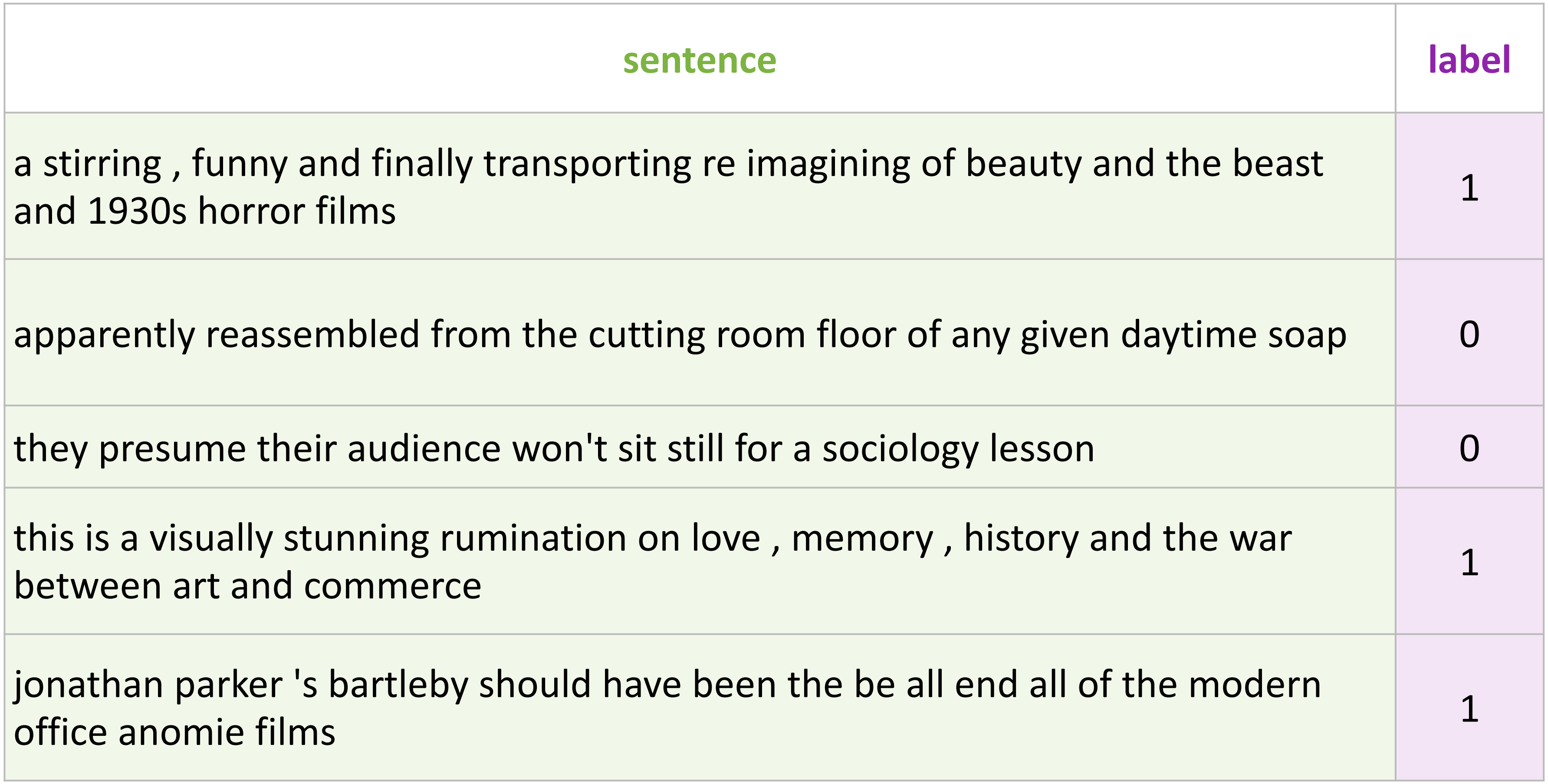
Source: Jay Alammar (2019), A Visual Guide to Using BERT for the First Time, http://jalammar.github.io/a-visual-guide-to-using-bert-for-the-first-time/

Score Benchmarks Logistic Regression Model on SST-2 Dataset

```
# Training
lr_clf = LogisticRegression()
lr_clf.fit(train_features, train_labels)

#Testing
lr_clf.score(test_features, test_labels)

# Accuracy: 81%
# Highest accuracy: 96.8%
# Fine-tuned DistilBERT: 90.7%
# Full size BERT model: 94.9%
```

Source: Jay Alammar (2019), A Visual Guide to Using BERT for the First Time, http://jalammar.github.io/a-visual-guide-to-using-bert-for-the-first-time/

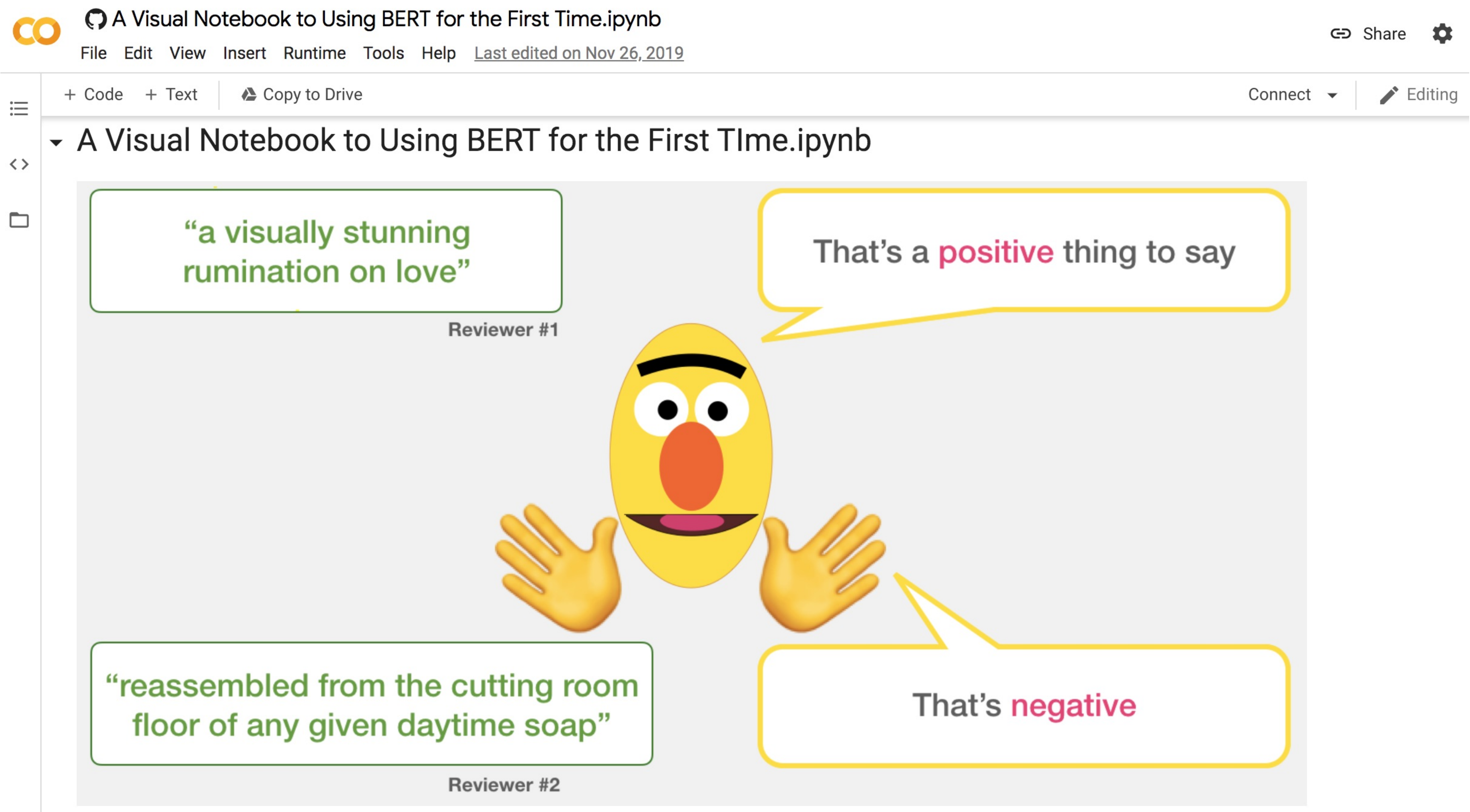
Sentiment Classification: SST2
Sentences from movie reviews



sentence	label
a stirring , funny and finally transporting re imagining of beauty and the beast and 1930s horror films	1
apparently reassembled from the cutting room floor of any given daytime soap	0
they presume their audience won't sit still for a sociology lesson	0
this is a visually stunning rumination on love , memory , history and the war between art and commerce	1
jonathan parker 's bartleby should have been the be all end all of the modern office anomie films	1

Source: Jay Alammar (2019), A Visual Guide to Using BERT for the First Time, http://jalammar.github.io/a-visual-guide-to-using-bert-for-the-first-time/

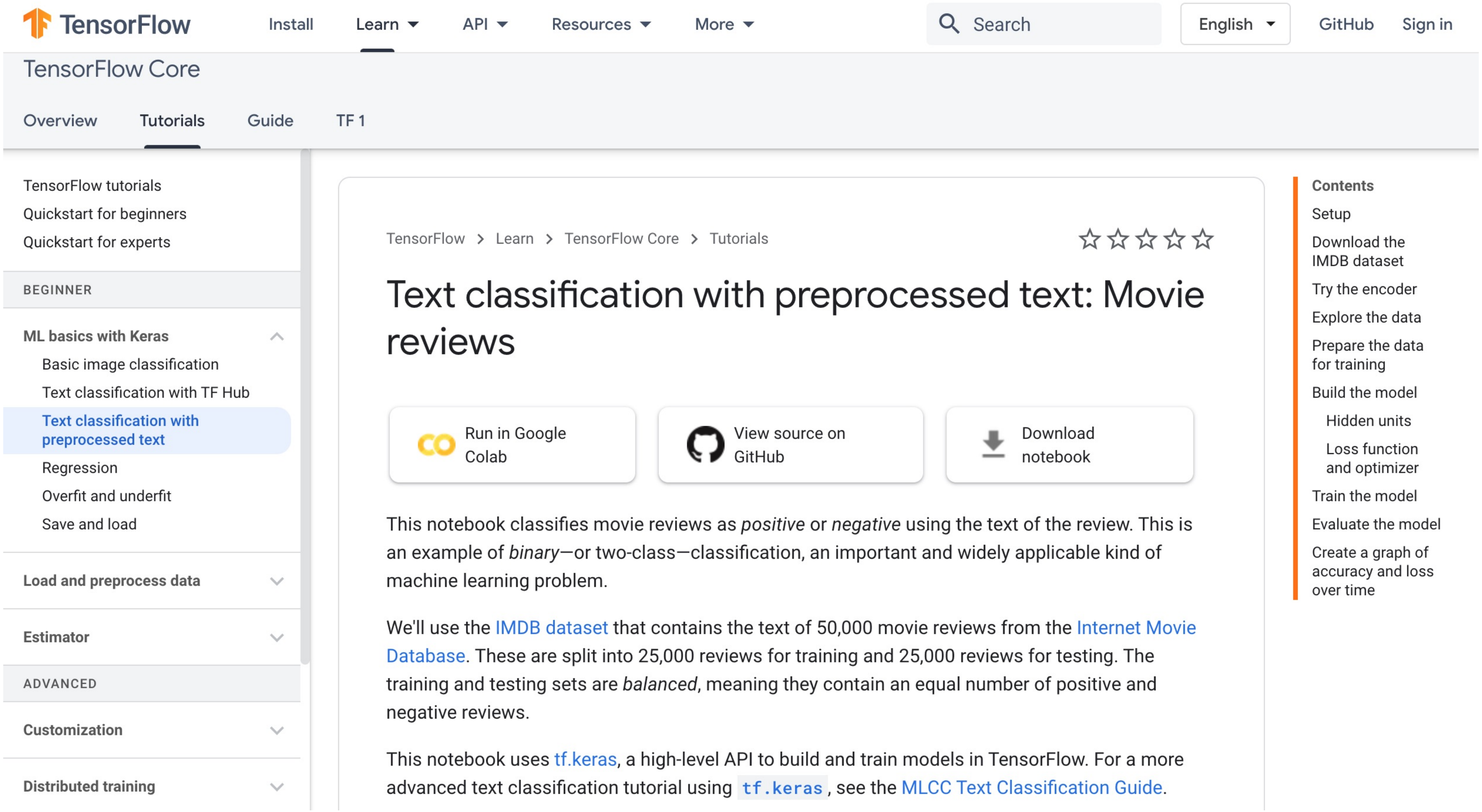
A Visual Notebook to Using BERT for the First Time



A Visual Notebook to Using BERT for the First Time.ipynb
File Edit View Insert Runtime Tools Help Last edited on Nov 26, 2019
Share
+ Code + Text Copy to Drive
Connect Editing
A Visual Notebook to Using BERT for the First TIme.ipynb
"a visually stunning rumination on love"
Reviewer #1
That's a positive thing to say
"reassembled from the cutting room floor of any given daytime soap"
Reviewer #2
That's negative

https://colab.research.google.com/github/jalammar/jalammar.github.io/blob/master/notebooks/bert/A_Visual_Notebook_to_Using_BERT_for_the_First_Time.ipynb

Text classification with preprocessed text: Movie reviews



TensorFlow Install Learn API Resources More Search English GitHub Sign in

TensorFlow Core

Overview Tutorials Guide TF 1

TensorFlow tutorials
Quickstart for beginners
Quickstart for experts

BEGINNER

ML basics with Keras
Basic image classification
Text classification with TF Hub
Text classification with preprocessed text
Regression
Overfit and underfit
Save and load

Load and preprocess data

Estimator

ADVANCED

Customization

Distributed training

TensorFlow > Learn > TensorFlow Core > Tutorials ☆☆☆☆☆

Text classification with preprocessed text: Movie reviews

Run in Google Colab | View source on GitHub | Download notebook

This notebook classifies movie reviews as *positive* or *negative* using the text of the review. This is an example of *binary*—or two-class—classification, an important and widely applicable kind of machine learning problem.

We'll use the IMDB dataset that contains the text of 50,000 movie reviews from the Internet Movie Database. These are split into 25,000 reviews for training and 25,000 reviews for testing. The training and testing sets are *balanced*, meaning they contain an equal number of positive and negative reviews.

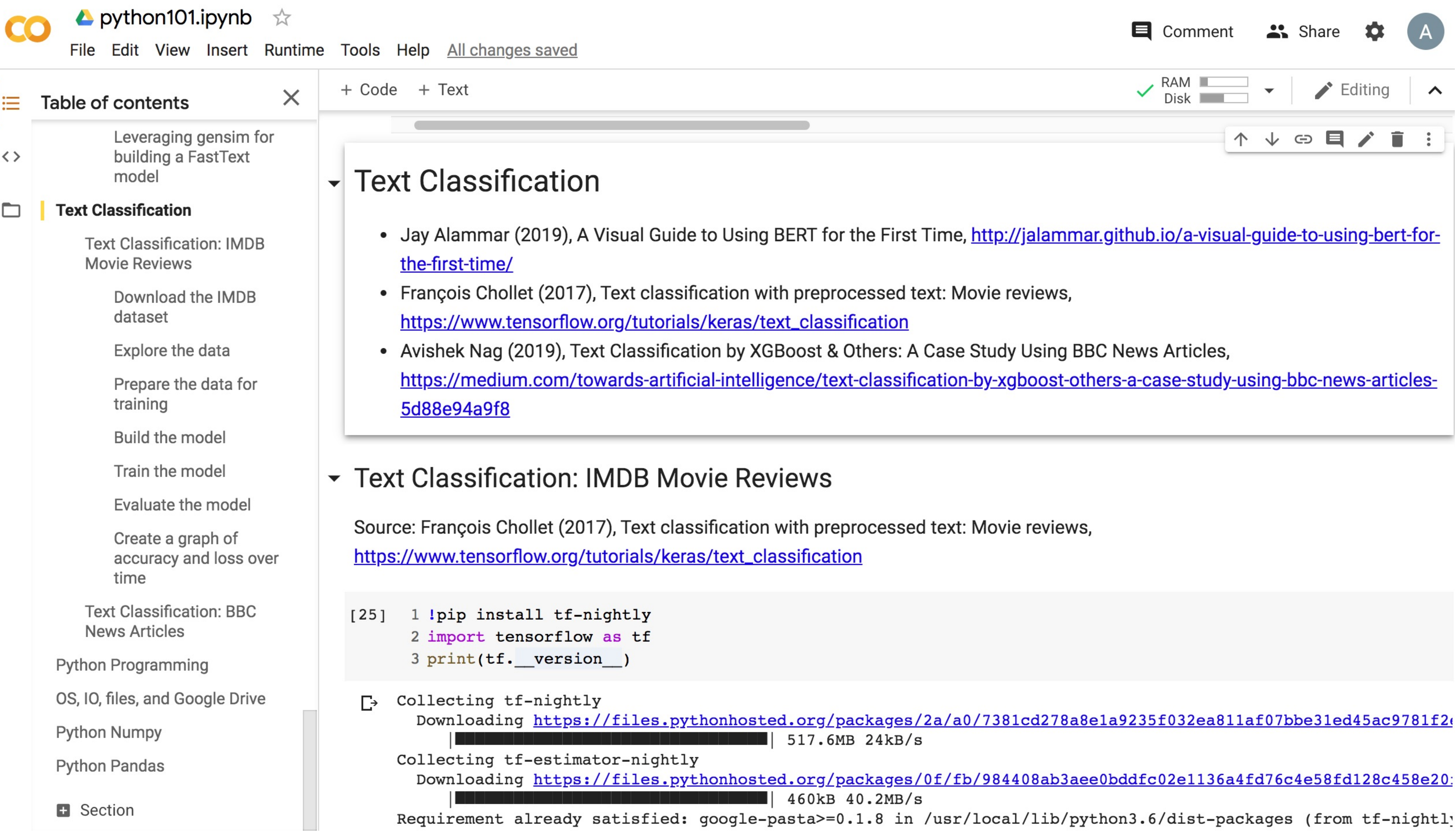
This notebook uses tf.keras, a high-level API to build and train models in TensorFlow. For a more advanced text classification tutorial using `tf.keras`, see the MLCC Text Classification Guide.

Contents
Setup
Download the IMDB dataset
Try the encoder
Explore the data
Prepare the data for training
Build the model
Hidden units
Loss function and optimizer
Train the model
Evaluate the model
Create a graph of accuracy and loss over time

https://www.tensorflow.org/tutorials/keras/text_classification

Python in Google Colab (Python101)

https://colab.research.google.com/drive/1FEG6DnGvwfUbeo4zJ1zTunjMqf2RkCrT



Text Classification

- Jay Alammar (2019), A Visual Guide to Using BERT for the First Time, http://jalammar.github.io/a-visual-guide-to-using-bert-for-the-first-time/
- François Chollet (2017), Text classification with preprocessed text: Movie reviews, https://www.tensorflow.org/tutorials/keras/text_classification
- Avishek Nag (2019), Text Classification by XGBoost & Others: A Case Study Using BBC News Articles, https://medium.com/towards-artificial-intelligence/text-classification-by-xgboost-others-a-case-study-using-bbc-news-articles-5d88e94a9f8

Text Classification: IMDB Movie Reviews

Source: François Chollet (2017), Text classification with preprocessed text: Movie reviews, https://www.tensorflow.org/tutorials/keras/text_classification

```
[25]  !pip install tf-nightly
      import tensorflow as tf
      print(tf.__version__)
```

```
Collecting tf-nightly
  Downloading https://files.pythonhosted.org/packages/2a/a0/7381cd278a8e1a9235f032ea811af07bbe31ed45ac9781f2
     |████████████████████████████████| 517.6MB 24kB/s
Collecting tf-estimator-nightly
  Downloading https://files.pythonhosted.org/packages/0f/fb/984408ab3aee0bddfc02e1136a4fd76c4e58fd128c458e20
     |████████████████████████████████| 460kB 40.2MB/s
Requirement already satisfied: google-pasta>=0.1.8 in /usr/local/lib/python3.6/dist-packages (from tf-nightl
```

https://tinyurl.com/imtkupython101

Python in Google Colab (Python101)

https://colab.research.google.com/drive/1FEG6DnGvwfUbeo4zJ1zTunjMqf2RkCrT

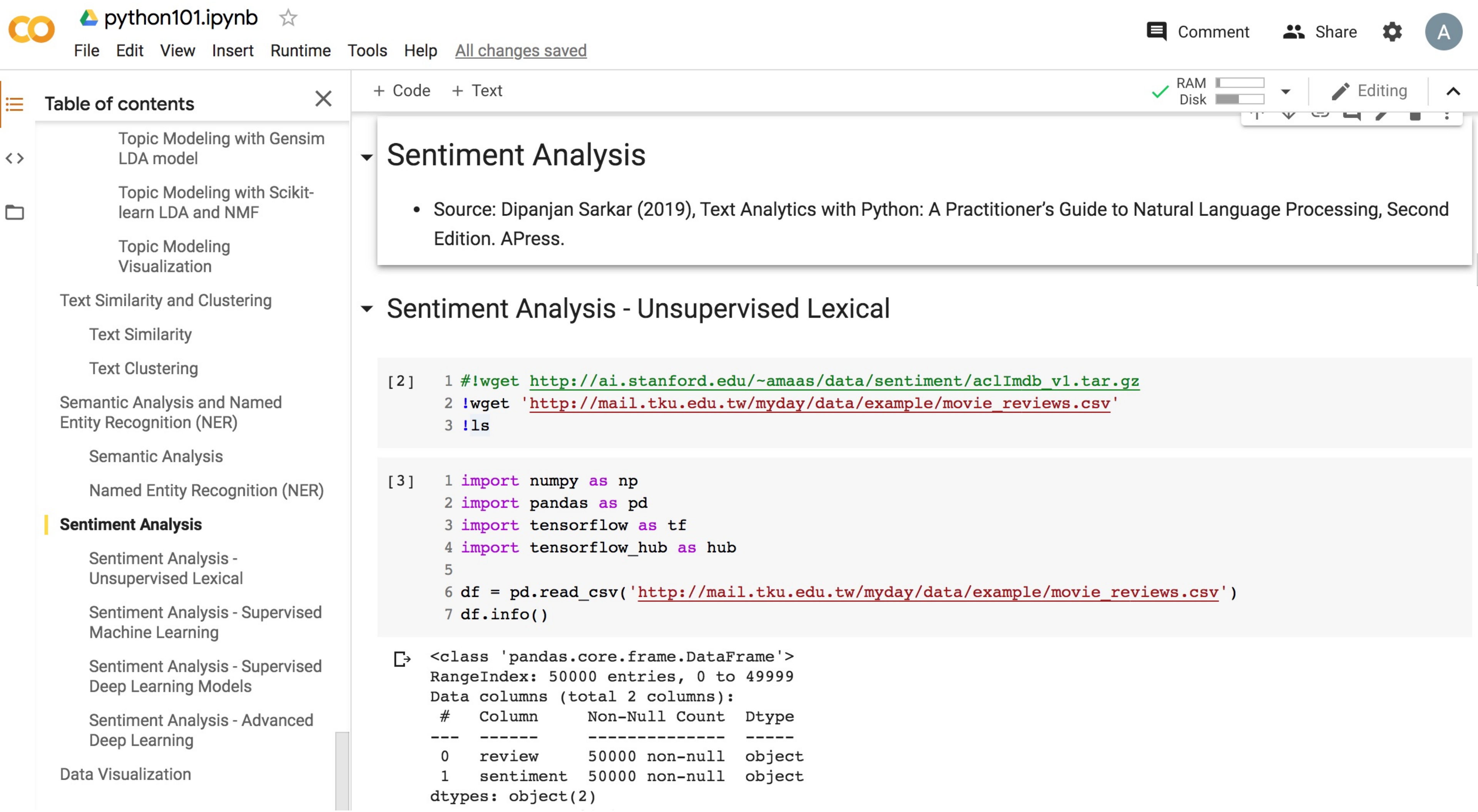


Table of contents

Sentiment Analysis

- Source: Dipanjan Sarkar (2019), Text Analytics with Python: A Practitioner's Guide to Natural Language Processing, Second Edition. APress.

Sentiment Analysis - Unsupervised Lexical

[2]

```
#!wget http://ai.stanford.edu/~amaas/data/sentiment/aclImdb_v1.tar.gz
!wget 'http://mail.tku.edu.tw/myday/data/example/movie_reviews.csv'
!ls
```

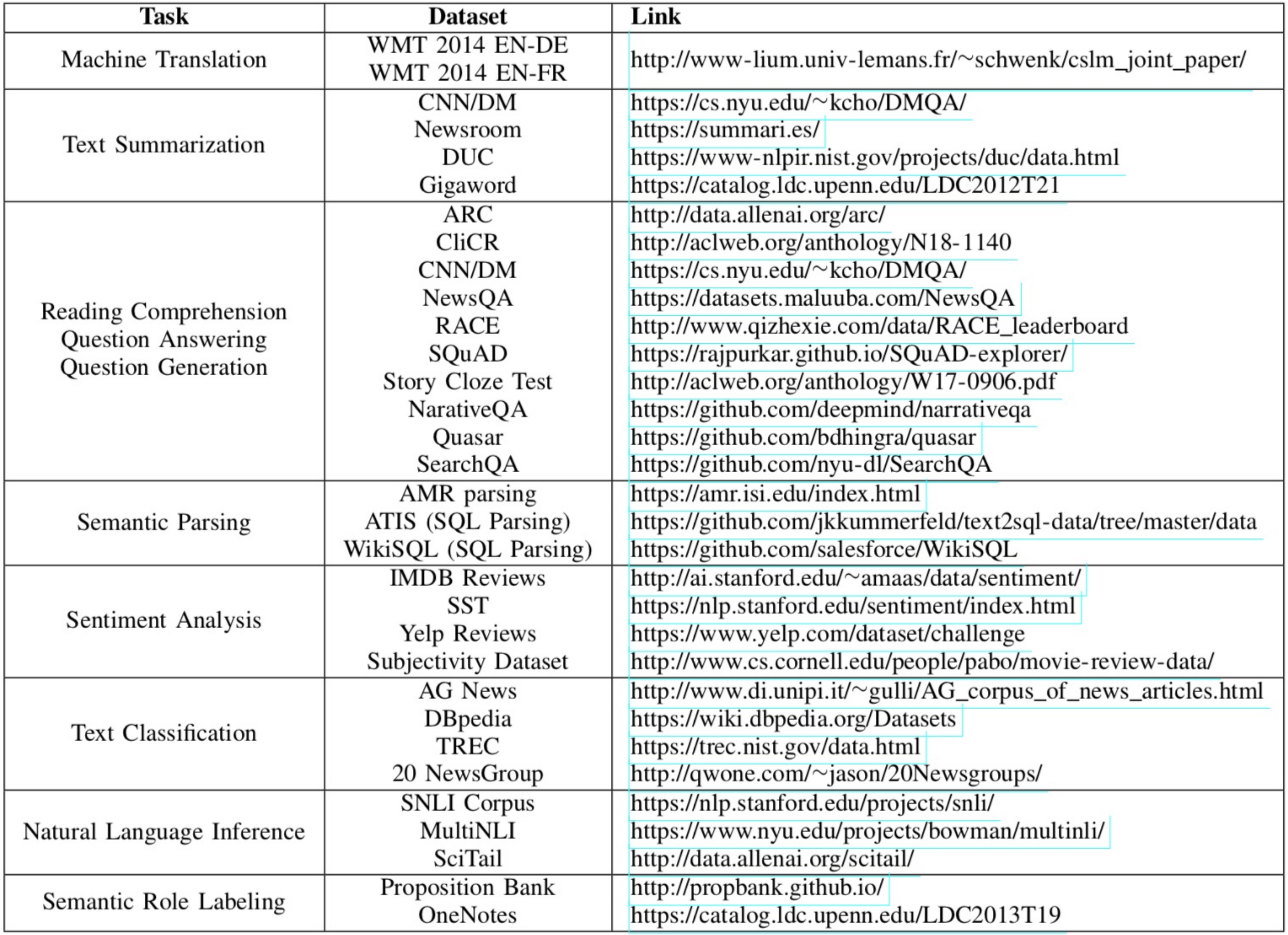
[3]

```
import numpy as np
import pandas as pd
import tensorflow as tf
import tensorflow_hub as hub

df = pd.read_csv('http://mail.tku.edu.tw/myday/data/example/movie_reviews.csv')
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 50000 entries, 0 to 49999
Data columns (total 2 columns):
 #   Column     Non-Null Count  Dtype 
---  ------     --------------  ----- 
 0   review     50000 non-null  object
 1   sentiment  50000 non-null  object
dtypes: object(2)
```

https://tinyurl.com/imtkupython101

NLP Benchmark Datasets



Task	Dataset	Link
Machine Translation	WMT 2014 EN-DE	http://www-lium.univ-lemans.fr/~schwenk/cslm_joint_paper/
	WMT 2014 EN-FR	
Text Summarization	CNN/DM	https://cs.nyu.edu/~kcho/DMQA/
	Newsroom	https://summari.es/
	DUC	https://www-nlpir.nist.gov/projects/duc/data.html
	Gigaword	https://catalog.ldc.upenn.edu/LDC2012T21
Reading Comprehension Question Answering Question Generation	ARC	http://data.allenai.org/arc/
	CliCR	http://aclweb.org/anthology/N18-1140
	CNN/DM	https://cs.nyu.edu/~kcho/DMQA/
	NewsQA	https://datasets.maluuba.com/NewsQA
	RACE	http://www.qizhexie.com/data/RACE_leaderboard
	SQuAD	https://rajpurkar.github.io/SQuAD-explorer/
	Story Cloze Test	http://aclweb.org/anthology/W17-0906.pdf
	NarativeQA	https://github.com/deepmind/narrativeqa
	Quasar	https://github.com/bdhingra/quasar
	SearchQA	https://github.com/nyu-dl/SearchQA
Semantic Parsing	AMR parsing	https://amr.isi.edu/index.html
	ATIS (SQL Parsing)	https://github.com/jkkummerfeld/text2sql-data/tree/master/data
	WikiSQL (SQL Parsing)	https://github.com/salesforce/WikiSQL
Sentiment Analysis	IMDB Reviews	http://ai.stanford.edu/~amaas/data/sentiment/
	SST	https://nlp.stanford.edu/sentiment/index.html
	Yelp Reviews	https://www.yelp.com/dataset/challenge
	Subjectivity Dataset	http://www.cs.cornell.edu/people/pabo/movie-review-data/
Text Classification	AG News	http://www.di.unipi.it/~gulli/AG_corpus_of_news_articles.html
	DBpedia	https://wiki.dbpedia.org/Datasets
	TREC	https://trec.nist.gov/data.html
	20 NewsGroup	http://qwone.com/~jason/20Newsgroups/
Natural Language Inference	SNLI Corpus	https://nlp.stanford.edu/projects/snli/
	MultiNLI	https://www.nyu.edu/projects/bowman/multinli/
	SciTail	http://data.allenai.org/scitail/
Semantic Role Labeling	Proposition Bank	http://propbank.github.io/
	OneNotes	https://catalog.ldc.upenn.edu/LDC2013T19

Source: Amirsina Torfi, Rouzbeh A. Shirvani, Yaser Keneshloo, Nader Tavvaf, and Edward A. Fox (2020). "Natural Language Processing Advancements By Deep Learning: A Survey." arXiv preprint arXiv:2003.01200.

Summary

- Unsupervised lexicon-based models
- Traditional supervised machine learning models
- Supervised deep learning models
- Advanced supervised deep learning models

References

- Dipanjan Sarkar (2019), Text Analytics with Python: A Practitioner’s Guide to Natural Language Processing, Second Edition. APress. https://github.com/Apress/text-analytics-w-python-2e
- Benjamin Bengfort, Rebecca Bilbro, and Tony Ojeda (2018), Applied Text Analysis with Python, O'Reilly Media. https://www.oreilly.com/library/view/applied-text-analysis/9781491963036/
- Kumar Ravi and Vadlamani Ravi (2015), "A survey on opinion mining and sentiment analysis: tasks, approaches and applications." Knowledge-Based Systems, 89, pp.14-46.
- HuggingFace (2020), Transformers Notebook, https://huggingface.co/transformers/notebooks.html
- The Super Duper NLP Repo, https://notebooks.quantumstat.com/
- Min-Yuh Day (2020), Python 101, https://tinyurl.com/imtkupython101