EPiC Series in Language and Linguistics
Volume 1, 2016, Pages 298–307
CILC2016. 8th International Conference on Corpus Linguistics

Extracting Domain-Specific Features for Sentiment Analysis Using Simple NLP Techniques: running shoes reviews

Antonio Moreno-Ortiz, Chantal Pérez-Hernández, Cristian Gómez-Pascual
Universidad de Málaga, Spain
{amo, mph, cgp}@uma.es

Abstract

This paper is a first attempt at designing a procedure to derive a domain-specific lexicon (both single words and multiword expressions) from an opinion corpus of specialized language. We use a corpus of reviews of running shoes as case study, compiled for this particular purpose. The main goal is to obtain a first approximation to the task of automatically extracting domain-specific expressions of sentiment to be used by our sentiment analysis software, Lingmotif.

1 Introduction

Sentiment Analysis (SA), also known as *opinion mining*, has received great attention in recent years, and has become of great interest not only for the NLP community, but also for linguists, thus becoming a significant sub-field within Computational Linguistics. According to Pang & Lee (2002), many other disciplines have also set the focus on human emotions and opinions, since they affect the way human beings communicate with each other, as well as the way they carry out an action. Even though machine learning techniques have produced reasonable results (Aue and Gamon, 2005), the fact that they are exclusively applicable to particular subject domains for which the algorithms must be trained is an obvious limitation.

For SA systems to have acceptable results, consumer product reviews, being primarily an opinion genre, have been regarded as a great source of sentiment-laden texts. From a practical perspective, most SA systems are actually focused on analyzing such texts, since they are seen as a means for companies to extract their users' opinions on their products and services. Clearly, the addition of Sentiment Analysis to social media measurement techniques has taken into consideration the various political and social content that can be found in the reviews of a product (Moreno-Ortiz and Pérez-Hernández, 2013).

Our approach to Sentiment Analysis is linguistically motivated, in the sense that it is based on searching the texts for lexical items that show some kind of semantic orientation, that is, items that are tagged in the SA application's database with a given valence. Similar systems are those described in Hatzivassioglou and McKewon (1997), Turney (2002) or Taboada et al. (2011).

Our system, Lingmotif, is currently under development and it is being implemented as a continuation of the work carried out in the creation of Sentitext (Moreno Ortiz et al., 2010, 2011). It implements the concept of Contextual Valence Shifters (CVS), as defined by Polanyi & Zaenen (2006). Lingmotif primary lexical resources include, therefore, a sentiment lexicon of individual words, a multiword expressions lexicon, and a set of contextual valence shifters that are applied to come up with a valence for given text segment. A thorough description of the application, however, falls outside the scope of this paper*. Suffice it to say that these lexical resources are capable of handling general-language texts, but fall short when dealing with specialized discourse (Moreno Ortiz et al 2011). Unlike Sentitext, Lingmotif is actually able to use multiple sets of lexical resources, thus tackling the domain-specificity issue: a given lexical item's valence may vary across differing domains. In Lingmotif, domain-specific lexicons may be added as plugins, and selected at runtime when a text belonging to that domain is analyzed. When a plugin lexicon is selected, the information contained in it will override the default, i.e., general-language, lexicon.

Obviously, domain-specific lexical resources need to be created before they can be used in the application. In what follows we describe a methodology to bootstrap this process from a domain-specific corpus.

2 Creating a domain-specific corpus

For this study we decided to use running shoes reviews as a case study, because this type of product is simple enough, in principle, to require little specialized knowledge to validate our results. In addition, the discourse features of product reviews are also well known, and follow a fairly simple structure. A review text may discuss the product as a whole in terms of a number of its defining features. Such features, however, may or may not be domain-specific. Table 1 below provides some examples of how certain features are applicable to certain products.

	comfortable	durable	fast	breathable
Car	✔	✔	✔	✘
Camera	✔	✔	✔	✘
Running shoes	✔	✔	✔	✔
Hiking boots	✔	✔	✘	✔

Table 1. Product features and domain specificity.

What is more, a product review will usually discuss the component parts of the item, which in turn have their own features. The different parts of a product and the way they are analyzed must also be taken into consideration to build a corpus for successful domain-specific Sentiment Analysis. Table 2 below exemplifies some of the component parts and relevant features taken into account when describing different products, including running shoes:

* Visit http://tecnolengua.uma.es/lingmotif for further details.

	camera	laptop	running shoe
Part	Lens	Keyboard	Midsole
Part	Screen	Screen	Upper
Part	Processor	Processor	Toe box
Feature	Weight	Weight	Weight
Feature	Speed	Speed	Speed
Feature	Low light capabilities	OS compatibility	Drop

Table 2. Component parts and their features.

2.1 Corpus composition

For this particular task, we decided to focus on reviews produced by professionals rather than end users, to avoid a number of issues associated to the latter type of texts, such as loss of focus, lack of specialized knowledge, spelling mistakes, bias, or use of non-technical jargon. Thus, our corpus was extracted exclusively from specialized running sites, most of them being written reviews, with some spoken reviews (video transcripts). Table 3 below shows the sources we used and the basic quantitative data of the corpus:

Raw text, no tags					**Lemmatized, no SW**		
Sources	reviews	tokens	types	T/T ratio	tokens	types	t/t r.
Runrepeat.com	399	867192	10246	1.18	476352	10102	2.12
Runningshoesguru.com	151	121926	6259	5.13	67172	4660	6.94
Runblogger.com	38	39917	3322	8.32	20557	2524	12.28
Gingerrunner.com	37	9983	1608	16.11	5587	1239	22.18
Irunfar.com	25	30467	3465	11.37	15531	2687	17.24
TOTAL	650	1,070,389	13,389	1.25	585,199	9,985	1.71

Table 3. Corpus composition.

2.2 Annotation

The corpus was annotated in XML format and an effort was made to preserve as much data as possible from the original (source, date, author, product, product type, product data, user ratings, review sections, etc.),, Obviously, different sources provided different data, the richest being *runrepeat.com*, which offered a very detailed set of data and content structure for every product:

```
<?xml version="1.0" encoding="UTF-8"?>
<review>
<productType></productType>
<productName> </productName>
<reviewSource><productRatings>
<runscore></runscore>
<userRating></userRating>
<userNumber></userNumber>
<expertRating></expertRating>
<expertNumber></expertNumber>
<fiveStar></fiveStar>
<fourStar></fourStar>
<threeStar></threeStar>
<twoStar>/twoStar>
<oneStar></oneStar>
</productRatings>
<productData>
<section name="areTheyForYou">
</section>
<section name="elements">
<elementOutsole>
</elementOutsole>
<elementMidsole>
</elementMidsole>
<elementUpper>
</elementUpper>
</section>
<section name="construction">
</section>
<section name="offers">
</section>
<section name="features">
<featureHeelCushioning>
</featureHeelCushioning>
```

<terrain> </terrain>	<featureForefootCushioning>
<archSupport> </archSupport>	</featureForefootCushioning>
<use> </use>	<featureStiffness>
<estimatedPrice></estimatedPrice>	</featureStiffness>
<estimatedPriceCurrency></estimatedPriceCurrency>	<featureStability>
<weight></weight>	</featureStability>
<weightUnit></weightUnit>	</section>
<drop></drop>	<section name="similarShoes">
<dropUnit></dropUnit>	</section>
</productData>	<section name="summary">
<reviewText>	</section>
<section name="intro">	</reviewText>
</section>	</review>
<section name="prosAndCons">	
<pros>	
</pros>	
<cons>	
</cons>	
</section>	

Table 4. XML annotation.

2.3 Lemmatization

We lemmatized the corpus, even though our aim was to use raw text as input, with a view to obtaining a simple procedure to extract sentiment cues from it. Our goal was to have a lemmatized version available for distribution and further research, and get a first approximation to what is offered by nouns and noun phrases for our next task (noun phrase chunking). Lemmatization was performed using AntConc (Laurence, 2014) and Someya's (1998) *e_lemma* (V.2).

The results of the lemmatization process revealed different language patterns related to running shoes. The most relevant patterns are the following:

1. Single nouns may point to product parts (e.g., *laces, tongue, midsole, outsole*) and product features (e.g., *cushioning, stability*, *drop*).
2. Proper noun sequences almost usually refer to brands and models, such as *Nike Pegasus, Saucony Triumph ISO* or *Hoka One One Clifton*, or other trademarks, such as *Asics Gel, Boa Closure System, Fluid Foam* or *Fulcrum Technology*.
3. Noun sequences and multi-word expressions are also used to identify either product parts (*toe box, heel counter* or *speed laces*), product features (*EVA foam, energy return* or *lug pattern*) or some other product-related characteristics (*racing flat, trail running* or *foot strike*)
4. Premodified noun phrases (single premodification), such as *abrasion resistant material* or *adaptive cushioning midsole,* which are included in the category of evaluative expressions.
5. Premodified noun phrases (coordinated premodification), where the polarity of the adjectives is unknown: *forgiving shoes*, *stiff but comfortable ride*.

3 Term extraction

3.1 Single Words

Single-word term candidates were extracted using AntConc's *Keywords* feature with the default log-likelihood method. A *keyness* value of 60 was found to be the optimal cut-off point in terms of signal-to-noise ratio.

An attempt to use the lemmatized version of the corpus made it clear that it was not a good idea. Relevant terms like *running* or *cushioning* (highly frequent terms) were lemmatized as *run* and *cushion*. Similarly, trademarks such as *Zoom* and *Boost* were assigned to their corresponding lemmas, which would have made it impossible to filter them in the next step.

This method returned 3,200 candidate keywords, which were then tagged for part of speech to come up with a list of 670 nouns. This list was manually filtered to obtain the final set of 248 single-word keyword nouns. Table 5 below shows the top 100 nouns sorted by frequency.

shoe	18969	experience	1137	strike	393	shape	207
heel	6676	ground	1125	movement	389	mud	201
version	3361	foam	1123	compound	383	rock	199
stability	3086	rating	963	gait	379	density	190
midsole	3027	size	922	absorption	377	series	189
height	2767	feature	834	responsiveness	371	layer	180
comfort	2295	stiffness	823	style	359	sockliner	177
feel	2059	track	790	structure	347	category	168
durability	1962	control	776	cushioning	327	resistance	166
pair	1952	shock	769	race	322	coverage	165
road	1942	arch	742	plate	286	versatility	162
performance	1795	grip	702	section	285	competition	154
drop	1790	eva	694	pattern	271	budget	152
technology	1691	transition	690	range	269	cycle	146
price	1681	breathability	578	racer	247	pace	141
ride	1472	market	554	sale	244	fitness	133
system	1435	energy	553	advantage	243	stack	132
protection	1394	distance	533	mileage	243	irritation	131
traction	1340	gravel	475	discomfort	224	package	119
runner	1328	quality	471	overpronation	220	portion	117
construction	1316	midfoot	455	brand	215	footwear	115
flexibility	1315	model	437	ventilation	215	slippage	113
terrain	1224	barefoot	427	moisture	212	rider	111
pronation	1191	line	420	abrasion	209	tag	110
speed	1168	addition	397	fabric	209	freedom	104

Table 5. Top 100 single nouns (sorted by frequency).

This process does have some limitations other than the manual filtering mentioned above. Although we obtained acceptable precision, recall was poor due to the limitations of the POS tagger (terms such as *upper, midsole, outsole* or *ride* were incorrectly tagged) or the commonness of some terms in general language, such as *tongue, sole, laces* or *weight*. Around 20 more terms were added after a new checking of the full list of keywords was carried out, since they might have been skipped during the POS tagging process.

Once all the keywords were extracted, we manually identified those nouns that referred to either product parts or product features, a relevant distinction for Sentiment Analysis. The result is shown in tables 6 and 7 below.

Parts (32)			
arch	lugs	platform	sole
evelet	mesh	quicklace	spike
fabric	midfoot	rand	strap
farefoot	midsole	rearfoot	toebox
gaiter	outsole	shoe	tongue
heel	overlays	sneaker	trainer

insole	package	sock-lines	tread
laces	plate	sockliner	upper

Table 6. Single noun product parts.

Features (79)						
absorbency	control	experience	<u>instability</u>	quality	speed	toughness
adaptability	coolness	feel	leverage	reflectivity	stability	traction
adjustability	coverage	firmness	lightness	reliability	stiffness	ventilation
affordability	craftsmanship	flex	longevity	resilience	structure	versatility
aggressiveness	curvature	flexing	materials	resiliency	style	visibility
agility	cushioning	fluidity	padding	resistance	supination	weight
balance	customization	functionality	performance	responsiveness	support	
Brand	design	geometry	plushness	rigidity	sustainability	
breathability	<u>discomfort</u>	grip	price	size	technology	
comfort	drainage	heaviness	proprioception	slippage	thickness	
comfortability	durability	height	propulsion	smoothness	tightness	
construction	elasticity	<u>imbalance</u>	protection	softness	torsion	

Table 7. Single noun features. Underlined words are negative words.

3.2 Multi-word Expressions

Multi-word Expressions (MWE's) play a significant role in specialized languages. Their management is thus critical for dealing with specific domains. Even though the scope of this paper is limited to extracting noun phrases, it is worth noting that most domain-specific MWEs come in this form; what is more, most of these actually come in a reduced set of possible syntactic patters (Arppe, 1995). Therefore, predicative statements (e.g. "We think the shoe was excellent") and adverb phrases (e.g. "The midsole performed well") were not searched for, even though they could be dealt with in a similar fashion.

Also, in keeping with the overall objective, i.e., to design a simple procedure for term extraction to be used in a Sentiment Analysis system, we designed simple a noun phrase chunker using NLTK (Bird et al. 2009). The following patterns were used:

- For product brands models and trademarks: **{<NNP><NNP>+}**
- For product parts and features: **{<NN|NNS><NN|NNS>+}**
- (Premodified) noun sequences, simple
  {<DT|PRP|PRP\$|POS>?<RB>?<JJ.*>+<NN.*>+}
- (Premodified) noun sequences, coordinated:
  {<DT|PRP|PRP\$|POS>?<RB>?<JJ.*>+<CC><JJ.*>+<NN.*>+}

3.2.1 Proper Noun Sequences

Using the sequence mentioned above, a total of 18,265 sets of two, three and four proper noun sequences were extracted. This, of course, included a large number of false positives and unique occurrences, which were removed with the use of regular expressions. Some other proper noun sequences (e.g. *Usain Bolt* or *Vancouver First Half Marathon*) were identified and manually added to or removed from the list, resulting in near 300 noun phrase sequences that were removed. A total of 2,000 relevant, domain-specific multi-word proper nouns were identified, mainly related to product models and brand trademarks, as shown in the table below.

Top 24 NNPs. sorted by frequency					
New Balance	1269	Mizuno Wave Inspire	70	Inov-8 Roclite	60
Puma Faas	136	Nike Free	70	Adidas Duramo	59
Brooks Ghost	92	Saucony Ride	67	Brooks Launch	59
Saucony Kinvara	88	Blown Rubber	66	Brooks Trance	59
Asics GT	84	Nike Lunarglide	66	Brooks Adrenaline GTS	58
Salomon XA Pro	84	Nike Zoom Terra Kiger	60	Road Stability Normal	57
Brooks Glycerin	74	Brooks PureConnect	60	Adidas Climacool Ride	55
Hoka One	74	Brooks Transcend		New Balance Fresh Foam	

Table 8. Top 24 NNs. sorted by frequency.

3.2.2 Common Noun Sequences

A total of over 30,000 sets of two, three and four common noun sequences were extracted by our NP chunker. As in the extraction of proper noun sequences, regular expressions were used to remove recurrent POS-tagging errors, since many adjectives ending in *-ing* were tagged as nouns and the third person singular of the verbs were tagged as the plural form of a noun (e.g. “lacks support” or “offers quality”). Some other errors were also identified and fixed, as “aberration resistance” or “trading pattern”.

Once the extracted list of sets was sorted out, a total of 7,200 unique noun sequences were removed and we kept those with more than two occurrences. The vast majority of these (87%) were sequences of two words.

Top 25 NNs. sorted by frequency			
running shoes	1402	speed work	234
running shoe	1316	stability shoe	199
heel height	777	cushioning system	195
heel cushioning	487	pronation control	194
forefoot cushioning	427	trail running	192
forefoot height	406	sock liner	191
toe box	356	foot motion	187
trail shoe	348	stiffness rating	181
lacing system	330	arch support	180
outer sole	255	trail shoes	180
running technique	154	racing shoe	168
shock absorption	237	heel strike	149
		performance shoes	148

Table 9. Most frequent common noun sequences

3.2.3 Premodified Noun Phrases

A total of 65,000 Noun Phrases premodified by an adjective were extracted from our corpus, using the following pattern:

{<DT|PRP|PRP\$|POS>?<RB>?<JJ.*>+<NN.*>+}

Determiners were removed in post-processing, so equivalent phrases like “a great fit” and “its great fit” were unified. In total, about 25,000 unique noun phrases were extracted.

4 Polarity Assignment

Once the domain-specific lexical items are identified and extracted, the next step involves assigning them a given polarity or valence. Polarity assignment was performed semi-automatically by leveraging on existing resources, i.e., the Lingmotif general-language polarity lexicon. A match query was performed between the list of extracted single words and adjectives present in noun phrases against the existing Lingmotif dictionary. When matched, the same valence was assigned automatically. Unmatched terms were checked manually. Figure 1 below summarizes the results obtained in this step.

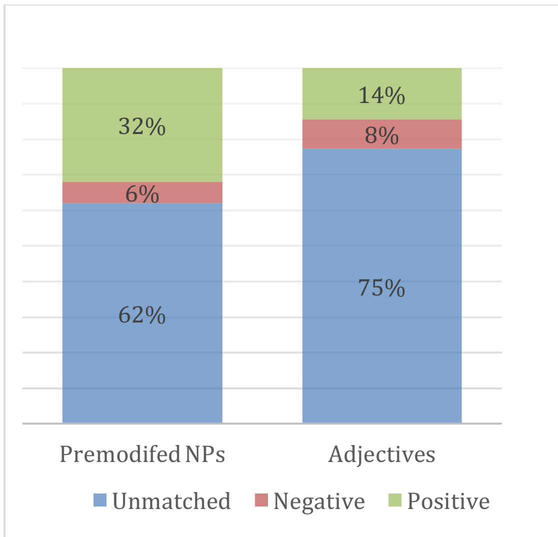

Figure 1. Results of matching adjectives against the Lingmotif database.

Unmatched premodifiers were, for the most part, compounds, but also misspellings (e.g. "abzorb") and trademarks (e.g., "adi-wear"). Table 10 shows the top 100 unmatched premodifiers. The superscript mark ($^{+}$) is used to indicate the polarity of the item (they all happen to be positive).

a-bound	advanced-level	all-terrain	analyze
abbreviated	adventure	all-time^{+}	anatomic
abound-based	aegis	all-weather	anatomical
above average^{+}	aerobic	all-white	anatomically-designed^{+}
above-mentioned	aerodynamic^{+}	allow	anatomically-engineered^{+}
abrasion-resistant^{+}	aesthetic^{+}	allowed	anatomically-placed^{+}
absolute	aesthetically-	almost-bare	and/or
absorb	appealing^{+}	almost-barefoot	anger
absorbs	aforementioned	almost-	angled
abzorb	age-long^{+}	featherweight^{+}	ankle
accelerated	ah-ha	almost-flat	anterior
accent	aha	almost-minimal	anterior/medial
accommodate	air-filled	almost-perfect^{+}	anti
accompanying	airflow	alongside	anti-abrasion^{+}
acetyl	airy	also-abrasion	anti-abrasive
achieve	aka	alternate	anti-bacterial^{+}

acorn-sized	albeit	alternative	anti-damage+
actual	all-black	altra	anti-debris+
add-on	all-day	altra-like	anti-friction+
added	all-important+	amateur	anti-microbial+
additional+	all-in-one+	american	anti-minimalist+
address	all-new	amphibious	anti-odor+
adi-wear	all-out	amplified	anti-pronating+
adjacent	all-purpose+	amply	antibacterial+
adjustment	all-round+	amply-cushioned+	antimicrobial+
	all-sewn+	anaerobic	

Table 10. Top 100 unmatched premodifiers.

5 Guessing Semantic Orientation from Coordinated Adjectives

A total of 1,733 coordinated adjectives acting as premodifiers of noun phrases were extracted. While dealing with coordinated adjectival structures, we can guess the orientation of the complete noun phrase when the semantic orientation of one of the adjectives is known.

Adjective + AND + Adjective: same semantic orientation:

“multi-purpose and durable outsole”

Adjective + BUT + Adjective: different semantic orientation:

“breathable but impermeable”

This approach to guessing orientation is not without its faults, though. There are more exceptions than desirable, where both of the adjectives are positive “small but durable traction knobs”.

6 Conclusions

As pointed out by several other studies, the expression of sentiment is domain-dependent to some extent. Therefore, a linguistically-motivated Sentiment Analysis system, such as Lingmotif, requires that lexical information for that particular domain be available. In this paper we have presented a relatively simple procedure to obtain such lexical resources from text corpora.

We have tried to automate the procedure as much as possible, but manual filtering was employed at certain steps, which of course is not optimal. Ideally, we should define an automated procedure that functions with as little user intervention as possible.

On the other hand, we have focused on noun phrases, which appear to carry most relevant information for us, but it remains to be seen whether other grammatical constructions (adjectives in predicative position, adverbs, verb patterns) would render relevant information not obtained by our NP approach.

References

Anthony, L. (2014). *AntConc (Version 3.4.3m)*. Tokyo, Japan: Waseda University. http://www.laurenceanthony.net/

Arppe, A. (1995). Term Extraction from Unrestricted Text. In *Proceedings of the 10th Nordic Conference of Computational Linguistics, Nodalida -95, Helsinki 29-30 May 1995*. University of Helsinki. http://www2.lingsoft.fi/doc/nptool/term-extraction.html

Aue, A., & Gamon, M. (2005). Customizing Sentiment Classifiers to New Domains: A Case Study. In *Proceddings of Recent Advances in Natural Language Processing (RANLP).* Borovets, Bulgaria.

Bird, S., Klein, E., & Loper, E. (2009). *Natural Language Processing with Python* (1 edition). Beijing ; Cambridge Mass.: O'Reilly Media.

Hatzivassiloglou, V., & McKeown, K. R. (1997). Predicting the semantic orientation of adjectives. In *Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics* (pp. 174–181). Madrid, Spain: Association for Computational Linguistics.

Jijkoun, V., de Rijke, M., & Weerkamp, W. (2010). Generating Focused Topic-specific Sentiment Lexicons. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics* (pp. 585–594). Stroudsburg, PA, USA: Association for Computational Linguistics. http://dl.acm.org/citation.cfm?id=1858681.1858741

Loper, S. B., Ewan Klein, Edward. (n.d.). *Natural Language Processing with Python*. http://shop.oreilly.com/product/9780596516499.do

Moreno-Ortiz, A. (2016). *Lingmotif*. Málaga, Spain: Universidad de Málaga. http://tecnolnegua.uma.es/lingmotif

Moreno-Ortiz, A., Pérez-Hernández, C., & Del-Olmo, M. (2013). Managing Multiword Expressions in a Lexicon-Based Sentiment Analysis System for Spanish. In *Proceedings of the 9th Workshop on Multiword Expressions MWE 2013* (pp. 1–10). Atlanta, Georgia, USA: The Association for Computational Linguistics. https://www.aclweb.org/anthology/W/W13/W13-1000.pdf

Moreno-Ortiz, A., Pérez-Hernández, C., & Hidalgo-García, R. (2011). Domain-neutral, Linguistically-motivated Sentiment Analysis: a performance evaluation. In *Actas del XXVII Congreso de la Sociedad Española para el Procesamiento del Lenguaje Natural* (pp. 847–856).

Moreno-Ortiz, A., Pineda Castillo, F., & Hidalgo García, R. (2010). Análisis de Valoraciones de Usuario de Hoteles con Sentitext: un sistema de análisis de sentimiento independiente del dominio. *Procesamiento de Lenguaje Natural*, *45*, 31–39.

Pang, B., & Lee, L. (2008). Opinion Mining and Sentiment Analysis. *Foundations and Trends® in Information Retrieval*, *2*(1–2), 1–135. http://doi.org/10.1561/1500000011

Polanyi, L., & Zaenen, A. (2006). Contextual Valence Shifters. In J. G. Shanahan, Y. Qu, & J. Wiebe (Eds.), *Computing Attitude and Affect in Text: Theory and Applications* (pp. 1–10). Dordrecht: Springer Netherlands. http://dx.doi.org/10.1007/1-4020-4102-0_1

Popescu, A.-M., & Etzioni, O. (2005). Extracting Product Features and Opinions from Reviews. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing* (pp. 339–346). Stroudsburg, PA, USA: Association for Computational Linguistics. http://doi.org/10.3115/1220575.1220618

Someya, Y. (1998). *e_lemma.txt*. Retrieved from http://www.lexically.net/downloads/e_lemma.zip

Taboada, M., Brooks, J., Tofiloski, M., Voll, K., & Stede, M. (2011). Lexicon-Based Methods for Sentiment Analysis. *Computational Linguistics*, *37*(2), 267–307.

Turney, P. D. (2002). Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)* (pp. 417–424). Philadelphia, USA.