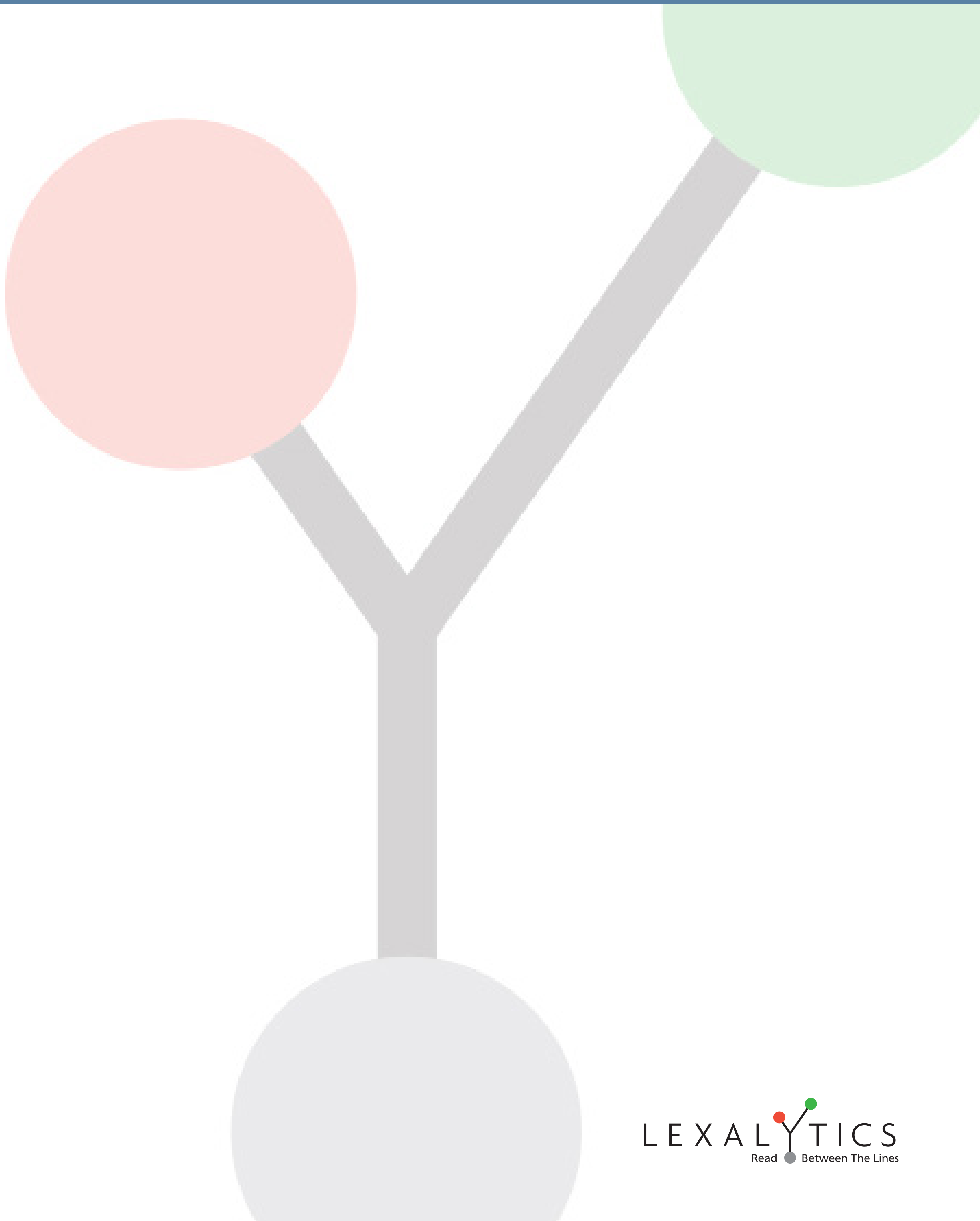

Entity Extraction

Whitepaper

LEXALYTICS
Read Between The Lines

AN INTRODUCTION TO ENTITY EXTRACTION

Text analytics is revolutionizing the way businesses approach the decision-making process. Never before has consumer feedback and public opinion been so easily interpreted on so massive a scale. At Lexalytics, we consider ourselves in the discovery business — and with our text mining solutions, you discover the who, what, and how of online consumer discussions:

- **Who's** talking
- **What** they're saying
- **How** they're feeling

These three categories are roughly definable by the three core functionalities of our software.

- Named entity extraction
- Themes, concepts, and facets
- Sentiment analysis

Many companies focus exclusively on the sentiment of their documents, paying little regard to who or what the sentiment is directed at. But let's get one thing straight: **sentiment is meaningless without context**. It's all well and good to have a vague sense of the sentiment directed at your brand, but true knowledge and informed decision-making is based in an understanding of **what sentiment** is being directed **where** and at **whom**.

The "who" of text analytics is called **entity extraction**. Our entity extraction methods are a cornerstone of the insights our text analytics solutions provide; Lexalytics has been in the text analytics industry for over a decade, and our software is powerful, refined, and customizable. The following whitepaper will show how named entities will inform your business.

We'll start with definitions.

WHAT IS AN ENTITY?

There are three important phrases to understand here: entity, entity extraction, and named entity extraction.

To begin, **entity extraction** is the process by which entities are identified from a block of text, and for our purposes this is synonymous with **named entity recognition**.

An **entity** in text, then, is a proper noun such as a person, place, or product. Lexalytics distinguishes proper nouns from generic nouns, which usually represent larger, vaguer concepts. So "Bill Gates", "Niagara Falls", and "iPhone" classify as entities, while "leader", "nature", and "technology" qualify as themes.

Lexalytics doesn't stop with common pronouns — our text mining tools can identify all of the following as entities:

- Companies
- Dates
- URLs
- Hashtags
- @Mentions (as in @lexalytics)
- Currency amounts
- Phone numbers

Even if it doesn't fit the traditional definition of a proper noun, anything treated as an entity in a document of text can be identified and tagged as such. That means the amount of money someone paid for a service they disliked, a popular hashtag in a collection of tweets, and so on. Several of our entity extraction systems even allow for full customization, so you can create your own definition of "entity".

POS TAGGING: THE FIRST STEP

Entity extraction is based on a technique called Part of Speech (PoS) tagging. Like the name implies, PoS tagging identifies the part of speech of any given word: noun, adjective, verb, adverb, etc.
PoS tagging for entity extraction focuses on proper nouns, which represent unique people, places, and things. Proper nouns are far more likely to be the entity focus of a document of text — that said, common (general) nouns can serve as entities in their own right. Lexalytics' software balances proper and common nouns to determine which are entities and which are more likely to represent themes.

ENTITY EXTRACTION: FIVE METHODS FOR THE RIGHT FIT

PoS tagging is the base from which our entity extraction methods leap: once tagging is complete, our systems offer a range of methods for identifying named entities. They are:

- Lists
- Patterns
- Regular Expressions
- CRF Model
- MaxEnt model

These methods range in complexity and have unique benefits and drawbacks.

Lists are just that: simple lists of named entities, like car manufacturers, people, or tree species. Lists are the most basic form of entity extraction. Once you've established a list of entities, the software pulls matches from the text. Both the beauty and disadvantage of list-based extraction is its simplicity: lists are clear-cut and easy to use and understand, but long lists are tedious to establish and list extraction only pulls direct references. Any tangential references, including pronouns, are overlooked.

As discussed earlier, P**art of Speech** (PoS) **patterns** are useful in determining entities. Noun phrases — phrases that involve a noun — in particular often represent entities. "Excellent soup" is a noun phrase (adjective-noun), as is "running dog" (verb-noun). We pick out these noun phrases and analyze them for likelihood of entity status. Verb phrases (such as "don't eat the cake") and other PoS patterns can and do represent entities, but noun phrases are more commonly entity-bearing.

Regular expressions allow you to define atypical named entities not included in the preset lists. A regular expression is essentially a search term for a specific item or type of item: gathering hashtags, @ mentions, and phone numbers all involve using regular expressions. If you want every phone number in a group of documents, for example, you might add search terms looking for the appropriate number patterns:

- (###)-###-####
- ##########
- ### ### ####
- Etc.

Searching for phone numbers is a great application of regular expressions. Of course, there are many different ways to write a phone number, and it will take time to add a search term for each variation — that said, it's quicker than training a new model. Regular expressions work well when searching for items that aren't necessarily unique but which follow a pattern (such as phone numbers). For very specific searches, your best bet is usually a list; for very general searches, our CRF model (below) is great. Regular expressions work at midlevel analysis, extracting entities that are vaguer than a concept but less specific than a pronoun.

The **Conditional Random Field** (CRF) **model** is a pre-trained system that automatically recognizes seven named entity types: Person, Place, Date, Company, Product, Job, and Title. Lexalytics hand-tagged entities of these types in a vast library of documents and fed them to our fledgling model, which analyzed the entities and learned from the patterns (including part of speech patterns). For example, our model learned that the phrase "works for" often precedes an entity (the name of a company). So when the phrase "works for" appears before a proper noun, the CRF model recognizes that the proper noun in question is the name of a company. Given enough of these clues (and we gave our model more than enough), the CRF model works with astonishing accuracy.

Lists and regular expressions allow for the definition of individual, unique entities and are great for smaller batches. But both systems require time to establish categories — and once defined, those categories are inflexible until you update them manually. That's why Lexalytics offers our **Maximum Entropy-based** (MaxEnt) **Model**. This toolset allows you to import and mark up your own training sets, to teach the computer yourself — allowing you to create entirely new categories of entity, like "Disease" or "Legal Term". Training a model in this way takes time and energy and is best used in specific circumstances, but when done correctly the results of a custom model can be well worth the investment.

HYBRID MODELS: OPTIMIZING FOR YOU

No single entity extraction method can serve every user's every need, but we know that our customers expect nothing but the best results from their applications of our software. That's why we've developed hybrid models, utilizing lists and rules to augment the Conditional Random Field system. The CRF model does the general work, and lists get down and dirty picking out the specifics you need.

Processing content about an election cycle, for example, is made easy with hybrid models. In this case, you're looking to pick up every mention of the politicians involved, but the total number of politicians is small — a perfect scenario for making a list. Enter names into a list, add the list as a modifier to the CRF model, and you'll guarantee that every name on the list will be reported as a person, regardless of their score in the CRF model.

Hybridizing our entity extraction system grants you, our customer, the wherewithal to fine-tune your results to your exacting standards.

SUMMARY

Lexalytics' suite of entity extraction tools lead the industry in their power and versatility. Our techniques give customers the flexibility to extract any type of entity using a range of tools — from simple lists of companies to highly sophisticated statistical models based on Part of Speech patterns. Once the entities have been gathered, we go that critical extra mile by assigning sentiment to each and revealing the context for each score, so that you're making the best-informed decisions you possibly can.

Lexalytics® is the industry leader in translating text into profitable decisions. Lexalytics deploys state-of-the-art on-premise and in-the-cloud text and sentiment analysis technologies that process billions of unstructured documents every day globally, transforming customers' thoughts and conversations into actionable insights. The on-premise Salience® and SaaS Semantria® platforms are implemented in a variety of industries for social media monitoring, reputation management and voice of the customer programs.

Lexalytics is based in Boston, MA, and has offices in the U.S. and Canada. For more information, please visit www.lexalytics.com, email sales@lexalytics.com or call 1-617-249-1049. Follow Lexalytics on Twitter, Facebook, and LinkedIn for updates and insights into the world of text mining.

LEXALYTICS
Read Between The Lines

320 Congress St
Boston, MA 02210

General Inquiries
1-800-377-8036

Sales
sales @lexalytics.com
1-800-377-8036 x1

International
1-617-249-1049