



## Yoogli Research White Paper

### Executive Summary

Vast quantities of new text data appear on the Web daily. Much of this text is produced by people commenting on their experiences with and impressions of various products and entities. These comments take the form of assertions about these products and entities, or attributes of the products or entities. The purpose of Yoogli's Research technology is to identify blocks of text that pertain to certain entities, and automate the classification of those blocks of text by user-requested type.

### Application

Yoogli Research helps brands listen, measure, and leverage consumer generated media (word of mouth) real-time. We help them understand the consumer emotion when interacting with their brand (positive or negative), supporting the data with verbatims, which aids in the understanding of the reasons for the brand perception. By trending this data Yoogli Research helps the brands understand the direction they are heading in consumer perception. This strengthens their ability to manage and control brand perception through product development and marketing. Currently, Yoogli Research has implemented assertion sets in field of health care. Consumer-generated buzz, at a low level, consists of assertions made by individuals about things. Yoogli Research core technology manipulates these assertions in very specific ways to quickly derive trends from sets of text documents. The remainder of this document describes how the technology performs this task.

### Assertions

An assertion is composed of three parts - an Entity, a Linkage, and a Vector. While full assertion quantification and comparison is not yet possible, it IS possible to assign a probability to the existence of a given assertion in some block of text. This section discusses how Yoogli's technology identifies and extracts assertions, as well as the operations performed on those assertions.

An Entity is a noun, compound noun, or assertion. Currently, nested assertions are not supported, as determining assertion boundaries is still an unsolved problem due to Vector complexity issues. Therefore, Yoogli's research technology concentrates its efforts on nouns and compound nouns. An example of a noun would be "fentanyl", which is a drug used in pain management. An example of a compound noun is "Merck's latest pain management drug". For our purposes, a compound noun can be any set of words not separated by a verb in the active voice or a paragraph break.

A Linkage is a word or set of words that tie an Entity to a Vector. Linkages almost always take the form of verbs or compound verbs, but can be terms of zero length (where adjectives of interest are used to describe Entities). The simplest non-zero linkage is the word "is". A Linkage can be several sentences. For our purposes, a Linkage is any set of words that are adjacent to an Entity, but do not contain another Entity or Vector.

A Vector is the heart of the assertion. It is the operation that is being performed on the Entity. It supplies us with the actual information that we seek - is the Entity neutral, good or bad? Does it relate to other Entities of interest? A Vector always contains an adjective. The potential complexity of a Vector is unlimited, and it is this characteristic which limits natural language processing technologies today. However, significant advantage can be obtained even with elementary Vector extraction, as most Vectors have a minimum core and that core is often composed of a single word or simple phrase. Entities can be linked to multiple Vectors. For example, the text "My Honda Accord is green. Unfortunately, this makes dirt easy to see." links the Entity "Honda Accord" with the Vectors "green" and "dirt easy to see".

### **Matching Assertions**

The primary job for Yoogli research is to assign a probability that a given block of text contains a given assertion. This is done by building assertions to match, then processing text against those assertions.

To build an assertion, it is necessary to specify an Entity and a Vector. The program automatically handles Linkages using some basic part-of-speech rules and precedence operations. Specifying Entities and Vectors is simply a matter of exhaustively enumerating every way to unambiguously refer to them. For example, the Entity "fentanyl" can be completely specified by listing the correct spelling of the drug, common abbreviations and misspellings, and indirect ways of referring to the drug such as brands like "Sublimaze" and "Duragesic". General words such as "drug", "pain medication", and "threshold pain" must also be linked to the primary Entity list. These secondary Entities are allowed to stand in for the primary Entity when a primary has already occurred in a block of text.

Specifying Vectors is similar to the process of specifying Entities. For example, the Vector "positive emotion" is composed of the words "good", "great", "best", and so forth. Obviously, more complex phrases (with sub terms that can occur in any order) are also important to tag as Vector elements. The system allows a limited use of regular expressions (counts and wildcards, to be exact) so that phrases like "<verb> well" can match large sets of non-Vectorized verbs, without applying when the word "well" is used as a noun or other type of word. Negation is also important to capture in this process, as a single negation can completely change the semantic direction of the Vector. Negation detection is built into the system - the user only needs to specify that a certain Vector is negatable (or has an opposite Vector).

Once the user has specified an Entity and a Vector, an assertion can be formed by combining them. For example, if the user wants to search a text repository for people talking positively about fentanyl, he must build a fentanyl Entity, a positive emotion Vector, and tie the two together.

### **Operations**

The system extracts assertions by first extracting Entities, then finding the boundaries of each Entity's influence within the text block, then extracting Vectors inside that area of influence, and finally matching those extracted

combinations to user-defined assertions. Finding the area of influence is the most important operation at this point, and the operation that determines the final accuracy of the query. The system relies on the user to provide a significant majority of the Entities that will be encountered in any query. The accuracy of the system smoothly increases with the number of Entities supplied. The reason for this is that the system determines Entity influence boundaries by extending them until a sentence that contains another Entity is reached, left to right. Where

Entities are not supplied, a Vector that occurs after the "dark Entity" will be incorrectly applied to the "live Entity" that has been identified. Since the set of Entities of interest comprises the whole of the language, Vectors are applied with decreasing probability the further they occur from the initial Entity. This allows the system to operate with some degree of confidence even when very few Entities are defined.

Experienced users will note that our approach is very imprecise with regard to individual occurrences of an assertion. However, when applied over significant quantities of text, this noise tends to cancel itself out and trends can be accurately measured. In this way, the Yoogli research technology is able to avoid the hard problem of complete language understanding, but is still able to leverage basic linguistic relationships to extract meaning over large quantities of text.

### **Semantics**

Semantic classification is an additional filter available to users of Yoogli Research. The system can categorize text based on semantic "attractors", which are densely populated and well-separated regions of semantic space. In this way, an Entity can be identified both by specific keywords and phrases as well as by the context in which it occurs.

### **Verbatims**

Once various blocks of text are classified according to the assertions loaded in the system, it is possible to automatically supply the best examples of text where the indicated assertions are satisfied. These segments of text, called verbatims, are used to support the general conclusions made by the automated scans where the research is being presented. A verbatim can be a paragraph, a set of paragraphs, or an entire page (where Entity density is particularly high).

### **Conclusion**

Yoogli Research provides a set of tools for users who need to perform assertion-related searches across large volumes of text. While the technology does not approach full understanding of language, it uses semantic operations to approximate this understanding with respect to the specific assertions provided by the user. As a result, users can quickly analyze sentiment trends, usage trends, and other widely scattered, non-structured data.