Executive Overview

Oracle’s social cloud text analytics platform is able to process unstructured text-based conversations to find the signal through its method of categorization and enrichment.

First, the platform defines posts as relevant or irrelevant to a category (“topic”) based on a combination of user defined key words and semantic “accept” or “reject” filters in context, instead of using key words alone.

Second, the platform enriches categorized posts by looking for snippets (the 1-2 sentences) within posts that semantically match to key indicators (consumer intentions, interests, or psychographics) such as purchase intent, problems, complaints, referrals, single moms, etc. The platform also scores snippets by sentiment, readability and subjectivity, and scores authors by demographics and Klout/influencer scores (where data is available). This enrichment process adds multiple and valuable meta-data to relevant posts and authors who made them, to support effective routing of posts within business process workflows, CX applications, and big data analytics and business intelligence reporting systems.

Because of this approach, the result from the platform is:

- Accurate – The platform cleanses and removes irrelevant posts that would be included if using keywords only, due to multiple meanings and/or ambiguous meaning of many key words;

- Comprehensive – The platform includes more relevant posts by avoiding the need to over filter as a method for excluding irrelevant posts;

- Insightful – The platform discovers new topics or dimensions that are relevant to a business through open-ended “white-space” analysis

- Clean – The platform cleans all incoming raw data to eliminate duplicate posts and SPAM. The platform’s cleaning algorithms have been refined and optimized using social media content
Oracle’s Blended Technology Listening Platform

Oracle has developed unique and proprietary IP around its text processing approach and specifically in its development and application of latent semantic analysis (LSA) as a method to automatically categorize and enrich text content in near real-time.

Brief Overview of Latent Semantic Analysis

Oracle’s LSA is an advanced form of statistical language modeling (SLM). LSA is a method for exposing latent contextual-meaning within a large body of text. It does this by looking at word usage (specifically, word co-occurrence) within a set of documents. Words, which appear in similar contexts, are assumed to have similar meaning and/or relational significance. LSA constructs a large matrix of term-document association data. Each cell in the matrix contains a weighted value, which is proportional to the number of times each term appears within each document in the set. The weights are structured such that rarer terms have greater weights. This allows more relevant terms to carry more weight to construct more accurate vectors of what/how consumers are talking about a category, brand, or product.

This matrix is then analyzed and reduced using singular value decomposition (SVD), a method of statistical transformation. SVD identifies relevant correlations between specific combinations of terms and documents. The process constructs document and term spaces for each term and document with an n-dimensional vector within the space. Following this process, each document occupies a specified position within the semantic space, with similar documents appearing near each other. Document-to-document similarity can then be performed by using simple vector distance measures such as cosine or Euclidean. Cosine is a trigonometric function that calculates the length of the adjacent sides (x and y-axis) of a hypotenuse. A Euclidean measure computes the square root of the sum of the squares of the differences between corresponding values deriving the shortest distance between two points that form a 90-degree angle on an x and y-axis.

Oracle combines its LSA technology with other forms of text processing, including:

- Keyword/Boolean search to assist in quickly initiating topic or dimension categorization queries;
- Natural language processing (NLP) to conduct speech analysis to further enrich snippet data through sentiment and readability scoring, voice of customer, and word usage analysis.
Oracle has implemented this approach, as compared to most other sophisticated vendors in the marketplace (who rely on combination of keyword/natural language processing), for the following reasons:

- **Cost of Client Setup** - For new domains, other vendors can spend a lot of time teaching the system the exact meaning of the vocabulary for each new area a client wants to build topics on. For example having to teach the system the difference between "drive" as a verb in an automotive context and "drive" as a piece of hardware in a computer context. Oracle’s latent semantic analysis captures this contextual difference automatically and on the fly, using semantic theming (also known as auto-clustering).

- **Processing Speed** - with social media content volumes reaching hundreds of millions of posts per day it is essential that content is processed as quickly as possible. Oracle’s Latent Semantic Analysis approach enables our system to operate at scale, categorizing hundreds of thousands of messages against tens of thousands of topics in minutes. These speeds can be difficult to achieve in systems that require interaction with extensive knowledge bases to determine a post’s meaning.

- **Data Accuracy** - Oracle’s categorization technique delivers exceptional categorization accuracy without sacrificing ease of topic setup or processing speed.

Oracle’s blended text processing approach is most suitable for clients who want a high degree of categorization accuracy, precise enrichment and automated scoring from noisy consumer conversations in an efficient and effective manner.

**Oracle’s Messaging Processing Platform Process**

Oracle’s text analytics approach, part of its Oracle Social Engagement and Monitoring Cloud Service, delivers its insights by consuming massive amounts of social conversation and then applying unique IP processes to extract meaningful, relevant information from
that mountain of data. With access to this information, Oracle’s customers are able to make intelligent business decisions and take more targeted engagement action.

Figure 2. Oracle Social Cloud Platform Text Analytics Message Processing Pipeline

Oracle’s message processing occurs across three stages:

1. Cleaning
2. Categorization
3. Enrichment
   - Snippetization
   - Dimension Assignment
   - Scoring

Stage 1: Cleaning

The first stage of message processing removes SPAM and duplicate posts from the incoming data stream.

De-duping

De-duping is accomplished in two ways: the Oracle system accepts only a single, unique instance of a URL for each post. This prevents the exact same social media message from entering the system as a result of a provider error or multiple providers collecting the same message.

When the URL is distinct but the post content is exactly the same, such as with multiple posts for blogs, press releases and boards, an algorithm identifies the duplicate content and prevents the post from entering the system. (Note: Oracle does not eliminate duplicate content in the case of tweets because re-tweets are considered to be relevant message volume because they come from unique authors.)

SPAM Removal
To remove SPAM, Oracle’s system uses a Bayesian SPAM filter that has been trained and refined with social media content.

Stage 2: Categorization

The second stage of message processing identifies and assigns relevant topics to each post. This process utilizes the keyword and semantic filters that have been defined for each topic by end users. Users create topics, starting with a simple key-word based query, and then using the Latent Semantic Analysis engine’s auto-clustering capability to choose which types of conversations to accept (green), and which types of conversations to reject (red). This straightforward process allows a user to achieve 95%+ categorization accuracy for both ambiguous and unambiguous search terms.

Figure 3. Diagram of User Selection of Accept and Reject Filters to achieve Signal Accuracy

This stage of processing is at the post level and produces a subset of relevant posts for each topic.

Stage 3: Enrichment

Snippetization/Indicator Assignment

The final stage of message processing enriches the relevant posts for each topic by attributing the posts with values for each assigned indicator. To achieve the best accuracy, this process occurs at the snippet level. A snippet is a subset of a post that is made up of the words in close proximity to user defined anchor terms (a.k.a. the key word(s)). Using a snippet rather than a full post results in a better indicator assessment because the words being evaluated are in close proximity to the anchor terms.

See Exhibit A for a Sample of Oracle’s Library of Indicators (semantic based)

Scoring

Following snippetization and dimension assignment, each snippet is scored for the following attributes:

- Sentiment – negative, neutral, positive
- Readability – complexity of writing style (0..100)
- Subjectivity – identification of individual expression (0..100)
For each author, scoring is provided for demographics when available (age, gender, and location) and Klout (influencer) score. These values are captured by our data aggregators and made available within our system.

Conclusion - Post Processed, Tagged Conversations & Authors

Oracle's Social Cloud Platform for Text Analytics can process large volumes of social conversations consistently, fast, at scale with topic/indicator assignment and other scoring meta-data for consumption in Oracle’s Social Engagement and Monitoring Cloud Service, or via a data feed for consumption within appropriate client systems.

Specifically, once configured the platform, the result is the ability to:

- Achieve pure signal by identifying social media posts as relevant to a specific business opportunity or problem—this is the topic creation/filtering part of CI user setup;
- Define how the posts are relevant in a way that leverages the content into actionable information—this is the enrichment/dimension-creation part of CI user setup.
Appendix A – Sample of Oracle’s Library of Indicators

The following are semantically trained indicators that are used as KPIs for specific industries. Additional ones can be created as required.

<table>
<thead>
<tr>
<th>INDUSTRY</th>
<th>Automotive</th>
<th>Travel/Hospitality</th>
<th>Retail</th>
<th>Financial</th>
<th>Telecom</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standard Indicators</td>
<td>Favorable</td>
<td>Favorable</td>
<td>Favorable</td>
<td>Favorable</td>
<td>Favorable</td>
</tr>
<tr>
<td></td>
<td>Unfavorable</td>
<td>Unfavorable</td>
<td>Unfavorable</td>
<td>Unfavorable</td>
<td>Unfavorable</td>
</tr>
<tr>
<td></td>
<td>Purchase Language</td>
<td>Customer Service</td>
<td>Purchase Language</td>
<td>Intent to Switch</td>
<td>Intent to Switch</td>
</tr>
<tr>
<td></td>
<td>Quality</td>
<td>Intent to Travel</td>
<td>Sales/Coupons</td>
<td>Credit Card</td>
<td>Customer Service</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>INDUSTRY</th>
<th>CPG</th>
<th>Media/Entertainment</th>
<th>Life Sciences</th>
<th>Technology/Hi Tech</th>
<th>Insurance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standard Indicators</td>
<td>Favorable</td>
<td>Favorable</td>
<td>Favorable</td>
<td>Favorable</td>
<td>Favorable</td>
</tr>
<tr>
<td></td>
<td>Unfavorable</td>
<td>Unfavorable</td>
<td>Unfavorable</td>
<td>Unfavorable</td>
<td>Unfavorable</td>
</tr>
<tr>
<td></td>
<td>Purchase Language</td>
<td>Watching</td>
<td>Safety</td>
<td>Quality</td>
<td>Intent to Switch</td>
</tr>
<tr>
<td></td>
<td>Sales/Coupons</td>
<td>Viewing Intent</td>
<td>Recall</td>
<td>Purchase Language</td>
<td>Purchase Language</td>
</tr>
<tr>
<td>Additional Available Indicators</td>
<td>Price, Advertising, Recall, Quality</td>
<td>Humor, Offensive, Trailers/Previews</td>
<td>Nutrition, Healthy Foods, Quality</td>
<td>Price, Customer Service</td>
<td>Auto insurance, Home insurance, Life insurance</td>
</tr>
</tbody>
</table>