Abstract
This whitepaper discusses the impact of building a sentiment analytics solution for content management systems. It highlights the importance of sentiment analysis, existing applications/use cases, the need to build such a solution and the benefits that can be extracted. It also shares best practices from Tavant’s CMS Center of Excellence with special focus on attitude analysis built using Apache Hadoop for CMS owners to analyse the sentiment of content on their platform.

Introduction
‘Sentiment’ can be defined as detection of attitude i.e. identifying disposition towards a subject. This subject can be a person, an object or a product. The metrics that characterize a sentiment are:

a. Source
b. Target
c. Type
d. Content

Sentiment analysis, in the context of this whitepaper, focuses on calculating the weighted polarity of the content (as positive, negative or neutral) and its strength in the text being processed. It also provides an edge over conventional methods by not being hypothetical or unnatural.

This information is useful for organizations to measure and evaluate the performance of their product or social campaigns. It allows marketers, brands and organizations to mine valuable information from the enormous data being generated every minute.

Need for Content Sentiment Analysis
The advancement in technological capability to achieve consistently accurate results in real-time is aiding companies to broaden use cases into a variety of domains. For example, consider the case of a socially-active content aggregator which publishes thousands of stories/articles on various subjects and topics. These articles are consumed by tens of thousands of users, who also share their feedback that could amount to millions of comments every day. This feedback is vital knowledge about customer preference and reaction to the news being disseminated. A method or mechanism to understand the opinions, emotions and attitudes of readers about different subjects, would provide the aggregator with consumer intelligence. This could in turn lead to better traction and, as a result, higher returns on investment. However, news aggregators may not have the technological capability to monitor this feedback and the reasons for this could include:
• The size of the data which cannot be managed manually
• Traditional data storage systems (generally tied to the CMS) which are not capable of processing data
• Lack of readily available plugins that provide for such capability
• Complexity of the text classification due to grammar, sentence structures, phrases and slang

**Understanding Sentiments**

To understand the sentiment behind a user comment requires the solution to behave like a human being to cull out intelligence. It needs to, among several other features, be able to

• Understand complex grammar
• Understand negations and positives in the same sentence. For example: “I would have liked it earlier, but now I hate it.”
• Distinguish between normal and extreme emotions. For example: “This is awesome” and “This is good.”
• Avoid misinterpretations due to combinations. For example: “I do not find it good.”

Developing a sentiment analytics solution for CMS can help solve this issue by analyzing the comments based on language rules. This can help to get more users to the portal and thus generate more revenue.

**The Three Stages in the Sentiment Analysis Process**

The ideal process of sentiment analysis should be an automatic workflow which follows a three-step approach:

**Data Collection**

The first step is the collection of comments from the content management system. We need to collect, clean and pre-process data as applicable to structure the same for analysis.

The sentiment analysis solution should be flexible and integrate easily with the backend content management system.

The data collection module should include the following components:

• A Cron to create logs with the required input data on a flat file
• Apache Flume to collect and aggregate the log file to HDFS (Hadoop Distributed File System) on a running Hadoop cluster
The Cron process writes the comments, along with the necessary metadata linked to the comments at a regular frequency which is configurable based on the volume and velocity of data for the CMS. These flat files are picked up by Apache Flume, and written to HDFS. Flume is also responsible for copying a snapshot of master data at regular intervals to ensure that HDFS is in sync with the current state of the CMS.

Data Analysis

In the second step, data should be processed to decipher desired information. The output of this step will be the sentiment type and its strength for user comments. With data available on HDFS, the next step is to run an analysis on this data.

For this step, we should run a bunch of options, leading to a choice between two processes:

- Batch Processing
- Real-Time Processing

Batch Processing:

For a batch process, we should use Apache Hive to churn the output.

Real-Time Processing:

For real-time processing with high-velocity data generation, we should use Storm.

Data Visualization

When information becomes available, the next step should be to visualize the data using charts and grids. The better the presentation, the better is the understanding of the results. These results should be presented as a dashboard and make for important intelligence for company stakeholders such as the management, strategists, and the sales and marketing teams. This visual and interactive information provides for better insights.

This data should be device-agnostic to be powerful. Analytics solutions, therefore, require responsive cross-platform presentation. Thus, ‘Attitude’ visualizations can be accessed not only on desktops but across devices such as smartphones and tablets.
Sentiment Analysis Solution

Content Engineering

By using a blend of emerging technology innovations, such as Big Data, Cloud, Mobile and Social systems, content engineering will take place and this will result in user experience being transported to an entirely new level. These innovations enable organizations to get strategic information from the content being generated on their systems by users worldwide as follows:

- Big Data integration allows a huge amount of data, generated by content management systems, to be analysed rapidly.
- Cloud deployment offers cost-effective infrastructure for running this analysis.
- Mobile-optimized interfaces designed as dashboards provide updated information with anytime-anywhere access.
- Social engagement is the basis for this concept as content is being generated on social media sites, in real-time and accessed globally by millions of users.
CMS solutions provide an effective way to monitor content polarity and understand the strength of peoples’ attitude. This enables gathering critical insights and planning content strategy based on the reaction to posted articles and information. It bridges the gap between modern day portals and analytics.

**Attitude Analysis**

The last step is the sentiment or attitude analysis, where a sentiment/attitude analytics solution can be plugged into a blog, portal or online journal running on any content management software. The aim of this sentiment or attitude analysis is to allow portals to measure and analyse the sentiment of the content on the same. It is an end-to-end framework of processes ranging from data gathering to data presentation.

This unique solution provides an opportunity to understand the attitude of the audience and is not limited to merely measuring their actions, such as likes, visits etc.

The attitude solution can be integrated with any CMS solution at the backend to get intuitive dashboards that present information, which can be viewed anywhere. It provides concrete answers to queries that could earlier only be guessed or speculated.

It employs more than one technology to improve the accuracy of sentiment identification and categorization.

To detect not only an emotion but also the intensity behind it, these steps should be followed:
• Apply Natural Language Processing to classify the text better.

• Use custom map-reduce algorithms to process the classified text against existing word banks which have been created to include not only straight formats of words, but also phrases and combinations encompassing a whole suite of scenarios.

• Then, with the help of statistical methods, a sentiment score is given to the classified text sentiment to come up with an overall score for the comment.

• All the comment scores are finally stored in a warehouse with the reference metadata, and this information can be accessed at multiple levels as per the business requirements.

Given below is an aerial view of the solution:

![Aerial View of the Solution]

Summary

To summarize, a content management portal can yield various strategic insights, for example, the best-loved author, best-liked sections, popular stories and most positively received articles, to name a few. Analytics is equally important, if not more, to record negative sentiments. Thus, by using sentiment analysis, business owners can better plan their customer-reach initiatives.
References:


About Tavant Technologies

Headquartered in Santa Clara, California, Tavant Technologies is a specialized software solutions & services provider that leverages its expertise to provide impactful results to its customers across North America, Europe, and Asia-Pacific. Having leveraged its unrivalled capabilities and domain insights to create game changing results for leading businesses across chosen industry micro-verticals, Tavant Technologies is known for long-lasting customer relationships, engineering excellence and passionate employees.

Founded in 2000, the Company employs over 1000 people and has been recognized as Top 25 Best Companies to Work For. www.tavant.com

About Author

Sarvesh Gupta helps design and implement analytics solutions for Tavant Technologies. He has extensive experience in big data and data science, while being actively involved in cutting-edge technology projects in the areas of big data mining, machine learning, and data visualization. He has exposure to multiple domains including advertising technology, broadcasting, gaming, and e-commerce. When not coding, he enjoys reading, painting, stock trading, movies and music.

Contact information

Tavant Technologies, Inc
3101 Jay Street, Suite 101
Santa Clara California 95054
United States

Tavant Technologies, Inc.
3101 Jay Street, Santa Clara, CA 95054, USA
Tel: (866) 9-TAVANT Fax: +1-408-519-5401