Sentiment Analysis of Digital Currency Discussions: A Machine Learning and Ontology Approaches

Atmane HADJI[1,*,†], Farid Boumaza[2,3,†] and Dina sirine Bali[4,†]

[1]LISI Laboratory, Computer Science Department, University Center A. Boussouf Mila, 43000 Mila, Algeria
[2]Computer Science Department, University of Mohamed El Bachir El Ibrahimi, Bordj Bou Arreridj 34030, Algeria
[3]LAPECI Laboratory , University of Oran1, Oran 31000, Algeria
[4]Department of Computer Science, University Center A . Boussouf Mila, 43000 Mila, Algeria

Abstract

A Sentiment analysis on social networks has become an increasingly important research field in recent years, driven by the rapid growth of social media and the vast amount of user-generated data. Understanding online opinions and sentiments is crucial for gaining insights into public attitudes and trends. In this study, we compare two approaches for sentiment detection: the first relies on ontologies, and the second utilizes machine learning techniques. Ontologies provide a structured framework to represent domain-specific knowledge, thus enhancing the accuracy of sentiment analysis. In the machine learning approach, we employed four algorithms: Support Vector Machines (SVM), K-Nearest Neighbors (K-NN), Decision Tree, and Random Forest. SVM demonstrated superior performance compared to other algorithms such as K-NN. Our approach was applied to sentiment analysis of Facebook discussions about Bitcoin, demonstrating the practical application of both ontology-based and machine learning techniques in the financial domain. The results highlight the effectiveness of both approaches in economic sentiment analysis, offering valuable insights into trends and sentiments that could be extended to other fields such as finance and commerce.

Keywords

Sentiment Analysis, Social Networks, Ontology, Bitcoin, Machine learning

1. Introduction

In recent years, social media has become a crucial platform where users share their opinions, sentiments, and experiences, creating an abundance of exploitable textual data. This surge in information has driven the need for sentiment analysis, a field dedicated to interpreting and categorizing the emotions and opinions expressed online. Sentiment analysis has applications in diverse areas such as marketing, finance, economics, and politics, where it enables the classification of opinions as positive, negative, or neutral. In the economic context, for instance, sentiment analysis helps to understand consumer and investor perceptions and to anticipate market trends.

However, accurately extracting opinions from vast quantities of textual data remains challenging. Traditional static indexing methods often fall short in their ability to capture the nuances and context in which sentiments are expressed. To address this, two approaches stand out in the literature: the ontology-based approach and the machine learning-based approach. The former utilizes a structured representation of domain knowledge, enabling each opinion to be associated with a specific semantic meaning, enhancing interpretability. The latter approach, on the other hand, relies on machine learning models that can automatically recognize the contexts in which opinions are expressed, offering improved precision through learning algorithms such as decision trees.

In this study, we present and compare these two methods for opinion extraction from online text, focusing on economic topics such as Bitcoin. On one hand, the ontological approach is examined for its

Proceedings of the International IAM'24: International Conference on Informatics and Applied Mathematics, December 04–05, 2024, Guelma, Algeria
*Corresponding author.
†These authors contributed equally.
a.hadji@centre-univ-mila.dz (A. HADJI); farid.pgia@gmail.com (F. Boumaza)
0000-0001-6706-6360 (A. HADJI); 0000-0002-9785-420X (F. Boumaza)

ability to provide precise semantic analysis. On the other, the machine learning approach is assessed for its capacity to recognize varied contexts automatically. This research aims to demonstrate the strengths and limitations of each method, offering insights into their applications for understanding economic trends and public perceptions in various domains.

2. Background and Related works

2.1. Rule-Based NLP

The extraction of Rule-based opinion extraction uses predefined patterns or guidelines to identify and extract subjective information, sentiments, or attitudes from text data. This approach is widely used in natural language processing (NLP) and sentiment analysis tasks. This approach relies on a set of predefined linguistic patterns, grammatical rules, or heuristics to process and analyze text data. These rules, designed by linguists or NLP experts, capture specific linguistic structures, sentiments, or entities within the text.

2.1.1. Subjectivity and Sentiment Analysis

Opinion extraction is a subtask of sentiment analysis, aiming to identify the sentiment or emotion expressed in a piece of text. Subjectivity refers to the extent to which a statement is influenced by personal feelings, opinions, or beliefs.

2.1.2. Key Components

The "Key Components" refer to the fundamental elements or essential techniques employed in the processes of opinion extraction and sentiment analysis. These components enable the detection, structuring, and interpretation of opinions expressed in texts ,they include:

- **Linguistic Patterns:** Rules are typically defined based on linguistic patterns, syntactic structures, or semantic cues, including specific keywords, parts of speech, or syntactic relationships that are indicative of opinions or sentiments.
- **Gazetteers:** A gazetteer is a list of words or phrases associated with specific categories or entities, used alongside rules to identify named entities or specific terms related to opinions.
- **Regular Expressions:** Regular expressions are powerful tools for defining complex patterns in text and can capture various linguistic features that indicate opinions.

2.2. Ontology-Based Approach

Ontology-based opinion extraction uses a structured, formal framework to represent domain knowledge, allowing for a more precise interpretation of opinions by linking opinion concepts and their relationships within an ontology. This method enhances the semantic understanding of text, enabling more contextual analysis of sentiments.

- **Semantic Representation:** The ontology provides a structure of concepts and relationships specific to the study domain, allowing each opinion to be linked to its semantic meaning. The concepts and relationships defined in the ontology help capture the implicit aspects of the expressed sentiments.
- **Knowledge Structure:** Unlike static rules, ontology represents a dynamic knowledge framework, allowing adaptation to context and language variations within opinions.
- **Opinion Modeling:** Opinions are integrated within the ontology structure, allowing them to be contextualized based on their relationships with other domain concepts, offering a more robust interpretation of the emotions and attitudes expressed.

2.3. Machine Learning-Based Approach

Machine learning-based opinion extraction uses trained models on large datasets to automatically identify sentiments and opinions in varied contexts. This approach adapts to language nuances without requiring predefined rules.

- **Automated Sentiment Classification:** Using supervised learning models like decision trees or neural networks, this method automatically categorizes opinions into positive, negative, or neutral sentiments.
- **Pattern Recognition:** Unlike static rule-based patterns, machine learning models detect complex patterns within text based on training data, capturing the nuances and subtleties of the expressed opinions.
- **Adaptability and Scalability:** Models can be retrained with new data to adjust to evolving trends or opinions, ensuring relevant sentiment extraction across diverse contexts.

This study explores and compares these two distinct methods ontology-based and machine learning-based to assess their effectiveness in opinion extraction, particularly in analyzing economic or social opinions expressed on social media. Each approach has unique strengths in terms of accuracy, semantic interpretation, and adaptability.

2.4. Related works

This section presents the state of the art in ontology-based and machine learning-based information extraction (IE) methods. Ontology-based IE methods leverage structured knowledge representations to capture complex relationships within specific domains. These approaches were initially inspired by semantic web technologies, using ontologies to represent hierarchical and interconnected knowledge structures. Ontology-based methods are widely applied in areas such as information retrieval and natural language processing, offering advantages in precise information categorization and supporting interoperability across systems. By defining specific entities and the relationships among them, ontology-based methods enable robust and contextually relevant information extraction that improves data consistency across applications.

Several studies illustrate the utility of ontology-based approaches for IE. For instance, an ontology-driven framework [1] leverages human expert knowledge to extract domain-specific information from unstructured text, adding structured information to a dedicated ontology. The system in [2] integrates AI with ontology creation to facilitate clinical data extraction, enabling medical practitioners to visualize patient information effectively. Another work, OntoHuman [3], introduces an automated ontology-based method to extract key-value pairs in the field of spatial engineering, allowing user feedback to refine ontologies and improve data extraction. Additionally, OBIESOF [4] is an ontology-based retrieval system for organic agriculture, structured to store and share agricultural knowledge, thus supporting future application development in this sector. A related study [5] applies an ontology-based system for land use analysis, integrating relevant geographical and legal criteria to enhance decision-making capabilities.

On the other hand, machine learning (ML)-based IE methods demonstrate significant flexibility and adaptability in processing unstructured data across various domains. Unlike rule-based systems, ML algorithms—such as Support Vector Machines, Random Forest, and deep learning models—identify patterns and extract relevant information by learning from large datasets, making them highly suitable for dynamic and diverse data sources. ML models have shown exceptional results in extracting structured information from complex data sources, including text, images, and documents.

Several studies highlight the efficacy of ML-based methods. A study on clinical data [6] used ML and NLP techniques to identify fracture types in radiology reports, showcasing the potential of ML for structured medical data extraction. Additionally, an information extraction system for clinical applications [7] demonstrates how ML can accurately capture contextual information from radiology reports, enhancing abnormality tracking. Another research [8] focused on ML-driven invoice processing,

where the LayoutLM model outperformed traditional methods in handling layout variations across unstructured invoices. In the domain of misinformation detection, [9] presented an ML-based approach for identifying COVID-19-related "fake news," leveraging medical features for enhanced detection accuracy. Moreover, recent works [10][11] demonstrated the effectiveness of transformer-based models in handling handwritten digital documents and complex resume data, illustrating how advanced ML models can transform unstructured data into usable knowledge. In summary, ontology-based and machine learning-based methods provide complementary strengths in information extraction. Ontologies offer structured, contextually relevant knowledge representation, while machine learning provides scalability and adaptability, especially in dynamic data environments. Together, these methods push the boundaries of information extraction, each bringing unique advantages to various applications and contributing to a richer understanding of domain-specific data.

3. Proposed Approach

The following architecture (Figure 1) depicts the detailed design of our opinion analysis system. The proposed system consists of several stages:

3.1. Data Collection

We get information from social network (Facebook) online. We processed comments related to fan opinions semi-automatically. We leverage the GATE platform (General Architecture for Text Engineering) to proficiently extract relevant comments from popular social media platforms such as Facebook and Twitter.

3.2. Pretreatment

In this step, we identified the comments related to the Champions League, then processed them in the next step. The filtering techniques applied to the corpus include more than one baseband. We filter the data by bypassing extra spaces and formatting elements to obtain plain text. Consequently, typos are corrected using automated and manual tools, and text normalization is followed, including the removal of special characters, spaces and punctuation.

Currently, social media worldwide is considered the most visited source for information on modern technologies like Bitcoin. Bitcoin is the most prominent cryptocurrency with the largest market capitalization. Additionally, it is a digital currency that users can only access online. Thus, online platforms play a crucial role in disseminating information to individuals about Bitcoin and how it is used. People mainly turn to social media when making purchase decisions, including buying or investing in Bitcoin, which is why we chose social media—specifically Facebook, as it gathers all segments of society.

In our study, we classified the factors influencing Bitcoin into three distinct categories: positive factors, negative factors, and neutral factors [12].

3.2.1. Positive Factors

We identified several positive factors impacting Bitcoin's increase in value, including but not limited to rising demand, institutional adoption, inflation and economic instability, heightened media coverage, and other elements.

3.2.2. Negative Factors

The depreciation of Bitcoin is influenced by multiple factors, some of which include high volatility, economic crises such as wars, high-interest rates, competition from other crypt ocurrencies, difficulty in using it as currency, and additional factors.

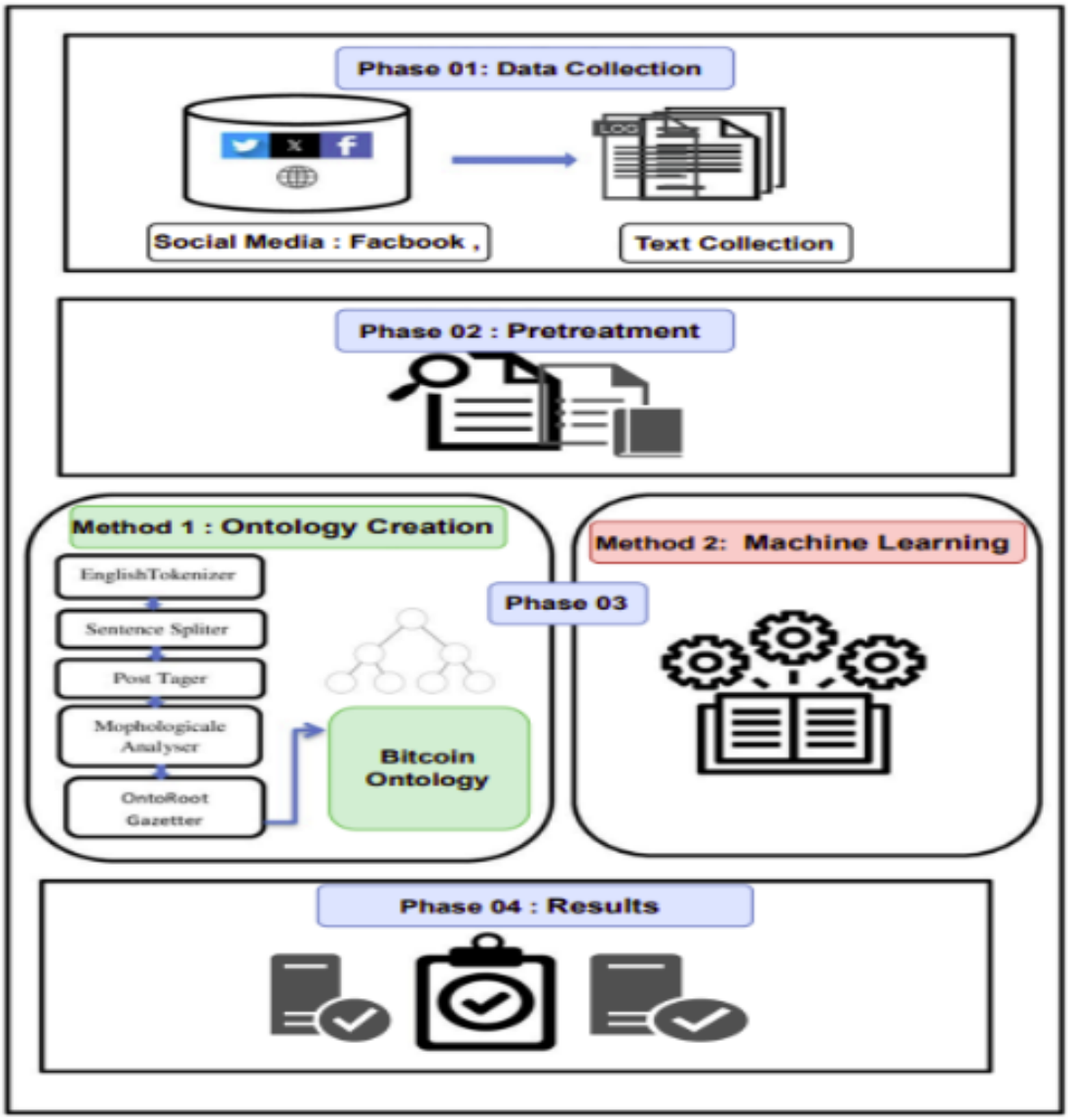

Figure 1: General architecture of the proposed system

3.2.3. Neutral Factors

There are also neutral elements, some of which are mentioned below: competition assessment, stability, and media updates.

The goal of extracting these factors that influence Bitcoin's value is to better understand the market and predict future trends, to enhance individuals' confidence in Bitcoin, encourage its usage, expand its application across different fields, improve the performance of exchanges and other platforms, and help more people understand this currency. Additionally, it aims to provide insight into the risks associated with investing in Bitcoin, protecting consumers from fraud.

We also focus on analyzing opinions about Bitcoin through posts and comments on Facebook regarding Bitcoin's price, satisfaction levels, and associated risks. Through this feedback, it is possible to:

- Determine the extent of Bitcoin's popularity;
- Assess whether people are optimistic or pessimistic about its future and better understand their needs;

- Measure public confidence in Bitcoin, their satisfaction level, and future expectations;
- Enable developers to design new technologies to improve market efficiency;
- Facilitate transactions and raise awareness of the risks associated with investing in Bitcoin, as well as provide insight into its influence on the economy and society.

3.3. Method 1 based Ontology

3.3.1. Ontology Creation Step

The flexibility of Ontology construction is a key aspect of this study. For this process, we adopted a top-down approach: starting with identifying high-level concepts, then refining them into more specific ones within our ontology, referred to as the "Bitcoin Ontology," which encapsulates the core knowledge of our work. This ontology was manually developed and then implemented in OWL format using the Protégé tool .

As outlined, the manual ontology development process involves the following steps [13]:

- Defining the domain and scope of the ontology;
- Considering the reuse of existing ontologies;
- Listing essential terms for the ontology;
- Defining classes and establishing the class hierarchy;
- Defining properties (slots) for the classes;
- Defining slot facets;
- Creating instances.

3.3.2. Tokenization

The Tokenizer divides text into simple words such as numbers, punctuation marks and many different types. For example, we have different words in Majestic and Minuscule, and among certain types of punctuation, etc. There is a "Token" annotation in the box, it should not be changed for different applications or text types.

3.3.3. Sentence Splitter

The sentence splitter is a cascade of finite-state transducers that segments text into sentences. This module is required for the tagger. The separator uses a list of gazetteer abbreviations to help distinguish phrase marking points from other types.

3.3.4. Part Of Speech Tagger

The tagger used is a modified version of the Brill tag, which assigns a part-of-speech tag to each word or symbol in the text. It is based on a lexicon and a set of default rules, which were learned from a large corpus from the Wall Street Journal. These elements can be adjusted manually if necessary.

Two additional lexicons are available: one for texts entirely in uppercase and the other for texts entirely in lowercase. To use them, simply load the appropriate lexicon, replacing the default one. In any case, the default rule set should always be used.

3.4. Metode 02 Machine Learning

Machine learning is a field of artificial intelligence that enables computer systems to learn and improve automatically from experience. By using algorithms and mathematical models, it analyzes data to recognize patterns and make decisions without being explicitly programmed. Machine learning applications are diverse, ranging from speech recognition and online product recommendations to fraud detection and autonomous driving. This field is rapidly advancing due to technological progress and

the increasing availability of massive datasets, opening new possibilities across many industrial and scientific sectors [14].

In this study, we investigate the application of machine learning techniques for opinion and sentiment extraction, leveraging four distinct algorithms: Support Vector Machines (SVM), K-Nearest Neighbors (K-NN), Random Forest Classifier, and Decision Tree Classifier. Each of these algorithms possesses unique characteristics and advantages, which significantly impact their effectiveness in identifying and extracting relevant information:

3.4.1. Support Vector Machines (SVM)

The Support Vector Machine (SVM) algorithm excels at classifying data by identifying the optimal hyperplane that maximally separates classes. In the realm of opinion and sentiment analysis, SVM is particularly effective for categorizing diverse types of information within complex textual data, ensuring precise and reliable classification.

For linearly separable data, the separation hyperplane can be determined by:

$$WTx + b = 0 \tag{1}$$

- $\mathbf{w}$ is the weight vector (or normal) of the hyperplane.
- $\mathbf{x}$ is the feature vector of a data point.
- b is the bias (offset) of the hyperplane.

3.4.2. Random Forest Classifier

The Random Forest algorithm improves classification performance by leveraging an ensemble of decision trees. By combining the outputs of multiple trees, it enhances generalization and reduces the risk of overfitting, making it particularly effective for managing diverse and noisy text data.

A Random Forest Classifier is an ensemble learning technique that merges the predictions of several decision trees to boost classification accuracy and mitigate overfitting. Each tree is trained on randomly selected subsets of data and features.

$$\hat{y} = \text{mode}\left(\{T_i(\mathbf{x}) \mid i = 1, 2, \ldots, N\}\right) \tag{2}$$

where:

- $p(y = 1 \mid x) = T(\mathbf{x})$ is the prediction of the i-th decision tree.
- N is the number of trees,
- The mode function returns the most common class label among all trees' predictions

3.4.3. K-Nearest Neighbors (K-NN)

The K-Nearest Neighbors (K-NN) algorithm is a straightforward yet powerful technique for classification and regression tasks. It classifies a data point by analyzing the majority class among its k-nearest neighbors in the feature space. This approach is especially advantageous for addressing multi-class problems and performs effectively when the data distribution is localized, making it a practical choice for various applications.

The K-Nearest Neighbors (K-NN) algorithm classifies a data point by measuring its distance to all other points in the dataset, selecting the k-closest neighbors, and assigning the class label most common among those neighbors. For a given data point x, the distance to each neighbor is computed using a metric like Euclidean distance:

$$d(\mathbf{x}, \mathbf{x}_i) = \sqrt{\sum_{j=1}^{n} (x_j - x_{i,j})^2} \tag{3}$$

where:

- $\mathbf{x}$ is the input feature vector.
- $\mathbf{x}_i$ is the feature vector of the i-th neighbor.
- n is the number of features.
- The class of $\mathbf{x}$ is determined by the majority vote among the k-nearest neighbors.

3.4.4. Decision Tree Classifier

Decision trees are highly interpretable models that operate by making a series of binary decisions. They are well-suited for extracting straightforward rules from textual data and provide clarity in understanding the criteria used for classification.

A Decision Tree Classifier divides data into subsets based on specific feature values, constructing a tree-like structure where each node corresponds to a decision guided by an attribute.

$$Gini(D) = 1 - \sum_{i=1}^{k} p_i^2 \tag{4}$$

where:

- k is the number of classes.
- p_i is the proportion of instances belonging to class i.

The tree continues to split until it reaches a stopping criterion, such as a maximum depth or minimum number of samples per leaf.

4. Results and Evaluation

After running the corpus with the use of JAPE and Gazetteer rules (figure 3), the system is now able to detect the entities named "Opinion Positive", "Opinion Negative" and "Opinion Neutral" corresponding to opinions on a Cryptocurrency ‘’Bitcoin”. Following the application of the Betcoin Opinion ontology to the corpus (Figure 2), the system can now identify named entities related to opinion of Betcoin.

The data used in the dataset for the first ontology-based method is the same as that used in the machine learning approach. This dataset is annotated with a range of attributes to support effective information extraction and sentiment analysis, including the classification of sentiments into Positive Opinions, Neutral Opinions, and Negative Opinions.

These annotations aim to evaluate machine learning models designed to extract relevant opinions related to Bitcoin. Figures 3 and 4 illustrate the results obtained for each algorithm used in our study: Support Vector Machines (SVM), K-Nearest Neighbors (K-NN), Decision Tree, and Random Forest.

To evaluate and compare the methods we studied, we will use metrics: Precision, Recall, and F-scale. Precision refers to the correctness of the retrieval, while recall refers to the completeness of the retrieval. The F-measure provides the harmonic mean between precision and recall [15].

According to [16] :

- Precision is the percentage of correctly recognized named entities (NE) among the recognized results:
$$\text{Precision} = \frac{\text{Number of correctly recognized NE}}{\text{Total number of recognized NE}} \tag{5}$$
- Recall is the percentage of correctly recognized named entities among the total entities that should have been recognized. It is a widely used measure in NLP evaluations:
$$\text{Recall} = \frac{\text{Number of correctly recognized NE}}{\text{Total number of NE in the corpus}} \tag{6}$$
- F-measure is the harmonic mean of precision and recall, providing a balanced evaluation:
$$F\text{-measure} = \frac{2 \cdot (\text{Precision} \times \text{Recall})}{\text{Precision} + \text{Recall}} \tag{7}$$

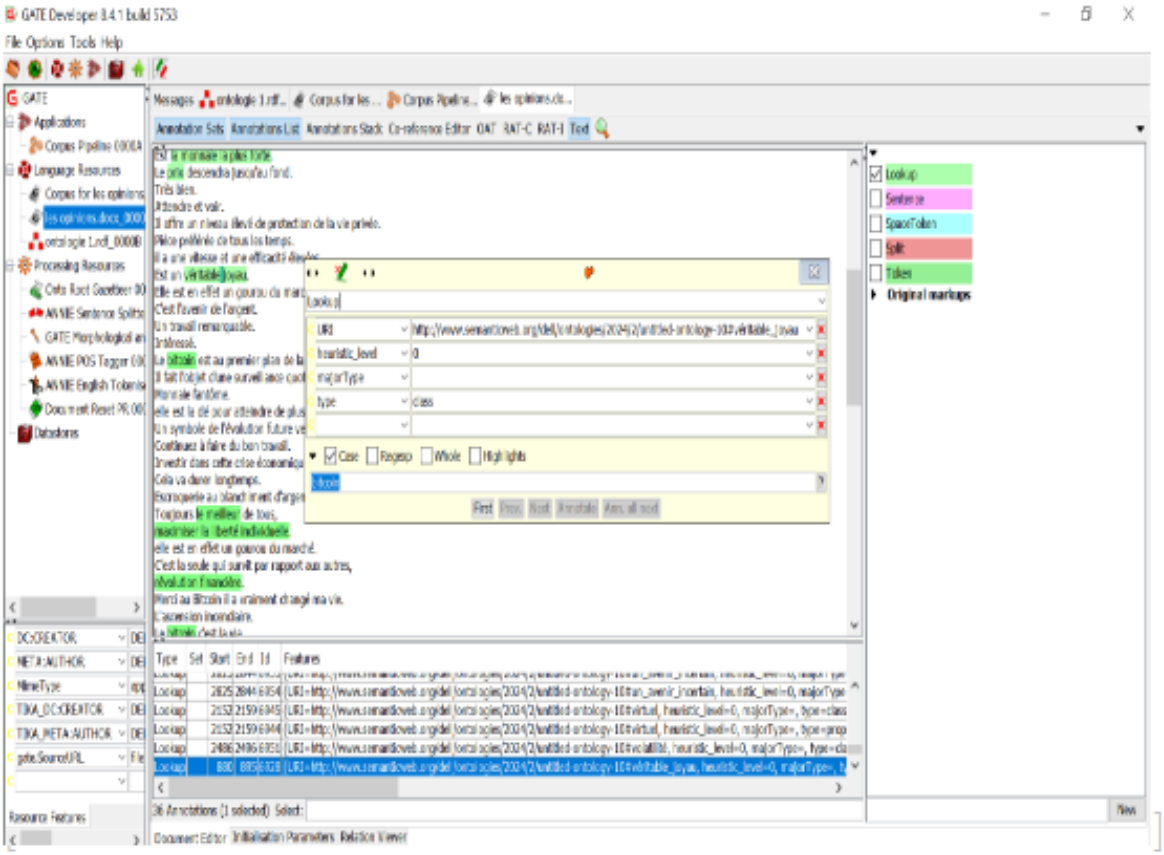

Figure 2: Results of Opinion Extraction in SVM and KNN Algorithms

Table 1
Results of Machine Learning (Average of four Algorithms)

Machine Learning	Precision	Recall	F-Measure
Negative Opinion	0.815	0.910	0.860
Neutral Opinion	0.882	0.9255	0.897
Positive Opinion	0.860	0.832	0.830
Total	0.880	0.860	0.850

5. Analysis and Discussion

5.1. Analysis and Discussion Machine learning

The results obtained in this study are highly satisfactory, as demonstrated by the Precision, Recall and F-mesure (see to Figure 3, Figure 4 and Table 1).

This section provides an in-depth analysis of the performance of the four algorithms (SVM, K-NN, Random Forest, and Decision Tree) used for opinion detection and sentiment analysis related to Bitcoin, based on data extracted from Facebook. The performance is compared in terms of precision, recall, and F-measure for three categories of opinions: negative, neutral, and positive.

5.1.1. Results Analysis

- **SVM:** The SVM classifier achieves the best overall performance, with an average precision of 0.90, a recall of 0.86, and an F-measure of 0.86, demonstrating its robustness in sentiment classification tasks. For negative opinions, the model exhibits strong detection capabilities, as evidenced by an

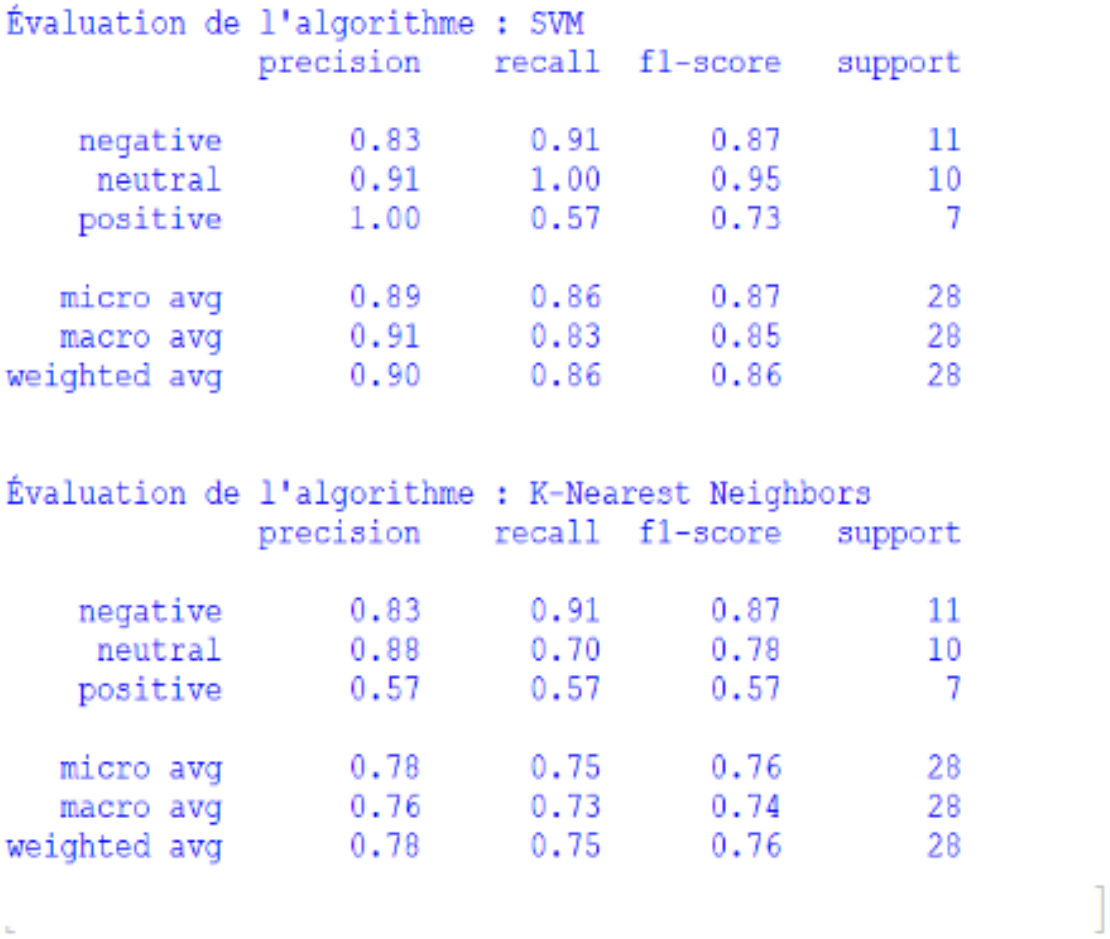

Figure 3: Results of Opinion Extraction in SVM and KNN Algorithms

F-measure of 0.87. Its performance is particularly remarkable for neutral opinions, achieving an exceptional F-measure of 0.95 and a perfect recall of 1.00, highlighting its ability to accurately identify and classify neutral sentiments. However, in the case of positive opinions, while precision reaches a flawless 1.00, the relatively low recall of 0.57 reduces the overall effectiveness in this category, resulting in an F-measure of 0.73.

- **K-Nearest Neighbors (K-NN):** The K-NN algorithm demonstrates the least effectiveness among the evaluated classifiers, with an average precision of 0.78, recall of 0.75, and F-measure of 0.76. Despite this, it performs reasonably well in detecting negative opinions, achieving an F-measure of 0.87, comparable to that of the SVM classifier. However, its performance declines notably for neutral opinions, where an F-measure of 0.78 is observed, primarily due to limited recall (0.70). The algorithm faces significant challenges in classifying positive opinions, as reflected in its particularly low F-measure of 0.57, highlighting difficulties in accurately capturing this sentiment category.
- **Random Forest:** The Random Forest algorithm delivers strong overall performance, achieving a precision of 0.88, recall of 0.86, and an F-measure of 0.85, underscoring its reliability in sentiment classification tasks. For negative opinions, it attains an F-measure of 0.83, which, although effective, is slightly lower compared to SVM and K-NN. Its performance in identifying neutral opinions is excellent, with an F-measure of 0.95, aligning closely with the results achieved by SVM. For positive opinions, the algorithm mirrors SVM's performance, achieving perfect precision (1.00) but exhibiting limited recall (0.57), leading to an overall F-measure of 0.73 in this category.
- **Decision Tree:** The Decision Tree algorithm demonstrates performance comparable to Random

Évaluation de l'algorithme : Random Forest

	precision	recall	f1-score	support
negative	0.77	0.91	0.83	11
neutral	0.91	1.00	0.95	10
positive	1.00	0.57	0.73	7
accuracy			0.86	28
macro avg	0.89	0.83	0.84	28
weighted avg	0.88	0.86	0.85	28

Évaluation de l'algorithme : Decision Tree

	precision	recall	f1-score	support
negative	0.83	0.91	0.87	11
neutral	0.83	1.00	0.91	10
positive	1.00	0.57	0.73	7
accuracy			0.86	28
macro avg	0.89	0.83	0.84	28
weighted avg	0.88	0.86	0.85	28

Figure 4: Results of Opinion Extraction in Decision Tree and Random Forest Algorithms

Forest, achieving an average precision of 0.88, recall of 0.86, and an F-measure of 0.85. For negative opinions, it performs on par with SVM, achieving an F-measure of 0.87, indicating strong detection capabilities. Its classification of neutral opinions is solid, with an F-measure of 0.91, although slightly below the performance of Random Forest and SVM. For positive opinions, similar to other algorithms, the Decision Tree achieves perfect precision (1.00), but its low recall (0.57) reduces the F-measure to 0.73, highlighting challenges in effectively capturing this sentiment category.

5.1.2. Comparative Discussion

- **Overall Performance:** SVM emerges as the top-performing algorithm, excelling in handling complex data and maximizing class separation, particularly for neutral and negative opinions. K-NN, despite its intuitive design, delivers the lowest overall performance, struggling notably with positive opinions due to its sensitivity to noise and limitations in capturing complex decision boundaries. Random Forest and Decision Tree display comparable performances, effectively capturing intricate patterns through their decision-tree-based methodologies. For neutral opinions, all algorithms, except K-NN, perform admirably. SVM and Random Forest stand out, achieving perfect recall (1.00), showcasing their precision in this category. However, detecting positive opinions poses a significant challenge across all models, with consistently low recall values (0.57). This difficulty may stem from data imbalance or the inherent ambiguity in distinguishing positive

Table 2
Results of Opinion Extraction (Ontology and ML Methods)

		Precision	Recall	F-measure
Positive	Method 1	0.560	0.740	0.630
Positive	Method 2	0.860	0.832	0.830
Neutral	Method 1	0.570	0.800	0.660
Neutral	Method 2	0.882	0.925	0.897
Negative	Method 1	0.620	0.850	0.710
Negative	Method 2	0.815	0.910	0.860

sentiments.
In terms of robustness and generalization, tree-based algorithms (Random Forest and Decision Tree) demonstrate strong resilience by mitigating overfitting risks. Despite this, they slightly trail behind SVM, which maintains the best overall performance in sentiment classification tasks.

5.2. Analysis and Discussion of the Tow Methods

This section presents a comparative analysis of two approaches used for sentiment analysis on Bitcoin-related posts from Facebook: Method 1 (Ontology-based) and Method 2 (Machine Learning-based). The results (See Table 2) are assessed based on three sentiment categories (Positive, Neutral, and Negative) and performance metrics: precision, recall, and F-measure.

5.2.1. Results Analysis

- **Positive Opinions:**
  Method 1: The F-measure of 0.63 reflects moderate performance in identifying positive sentiments, limited by lower precision (0.56).
  Method 2: With an F-measure of 0.83, Method 2 significantly outperforms Method 1, driven by high precision (0.86) and balanced recall (0.832).
- **Neutral Opinions:**
  Method 1: Achieves an F-measure of 0.66, with good recall (0.80) but relatively low precision (0.57).
  Method 2: Excels in detecting neutral opinions, achieving an F-measure of 0.897, the highest among all categories. This is due to strong precision (0.882) and near-perfect recall (0.925).
- **Negative Opinions:**
  Method 1: Demonstrates acceptable performance with an F-measure of 0.71, supported by recall (0.85) and moderate precision (0.62).
  Method 2: Outperforms Method 1 with an F-measure of 0.86, indicating better reliability in detecting negative sentiments, with precision (0.815) and recall (0.91) both being strong.

5.2.2. Comparative Discussion

- **Overall Performance:**
  Method 1: while demonstrating moderate performance, relies heavily on predefined rules and domain knowledge, limiting its flexibility and adaptability to nuanced language variations in social media posts.
  Method 2: (Machine Learning-based) consistently outperforms Method 1 (Ontology-based) across all sentiment categories. This is largely due to its ability to learn complex patterns in data and generalize well to unseen examples.
- **Neutral Opinions:**
  Method 1: exhibits higher recall values across all categories compared to its precision, suggesting

a tendency to detect more instances (including false positives).
Method 2: in contrast, achieves a better balance between precision and recall, reducing false positives while maintaining strong detection rates.

6. Conclusion

This study has provided an in-depth evaluation and comparison of ontology-based and machine learning-based approaches for sentiment analysis of Bitcoin-related discussions on social media, specifically Facebook. The results indicate that machine learning algorithms, particularly SVM, outperform both other algorithms (such as K-NN) and the ontology-based method in terms of precision, recall, and F-measure. While the ontology-based approach offers value through domain-specific knowledge representation, it falls short in flexibility and overall performance.

The strength of machine learning lies in its adaptability to complex and heterogeneous data, whereas ontologies provide a structured framework for capturing semantic relationships. These complementary attributes highlight the potential of hybrid approaches that combine the strengths of both methodologies.

Future research could explore hybrid methods to enhance both accuracy and interpretability. Incorporating additional datasets from diverse social media platforms and employing techniques such as data rebalancing may help address biases in certain sentiment categories, particularly positive opinions. Additionally, advanced deep learning models like BERT or GPT could further improve sentiment analysis by capturing the nuanced linguistic contexts of social media discussions. Expanding these methodologies to other domains, such as economics or healthcare, could open up new avenues for sentiment analysis applications.

Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

References

[1] R. Anantharangachar, S. Ramani, S. Rajagopalan, Ontology guided information extraction from unstructured text, arXiv preprint arXiv:1302.1335 (2013).

[2] S. Jusoh, A. Awajan, N. Obeid, The use of ontology in clinical information extraction, in: Journal of Physics: Conference Series, volume 1529, IOP Publishing, 2020, p. 052083.

[3] K. Opasjumruskit, S. Böning, S. Schindler, D. Peters, Ontohuman: ontology-based information extraction tools with human-in-the-loop interaction, in: International Conference on Cooperative Design, Visualization and Engineering, Springer, 2022, pp. 68–74.

[4] A. A. Abayomi-Alli, S. Misra, M. O. Akala, A. M. Ikotun, B. A. Ojokoh, et al., An ontology-based information extraction system for organic farming, International Journal on Semantic Web and Information Systems (IJSWIS) 17 (2021) 79–99.

[5] M. Al-Ageili, M. Mouhoub, An ontology-based information extraction system for residential land-use suitability analysis, International Journal of Software Engineering and Knowledge Engineering 32 (2022) 1019–1042.

[6] J. Fiebeck, H. Laser, H. B. Winther, S. Gerbel, Leaving no stone unturned: using machine learning based approaches for information extraction from full texts of a research data warehouse, in: International Conference on Data Integration in the Life Sciences, Springer, 2018, pp. 50–58.

[7] J. M. Steinkamp, C. Chambers, D. Lalevic, H. M. Zafar, T. S. Cook, Toward complete structured information extraction from radiology reports using machine learning, Journal of digital imaging 32 (2019) 554–564.

[8] F. Krieger, P. Drews, B. Funk, Automated invoice processing: Machine learning-based information extraction for long tail suppliers, Intelligent Systems with Applications 20 (2023) 200285.

[9] F. Fifita, J. Smith, M. B. Hanzsek-Brill, X. Li, M. Zhou, Machine learning-based identifications of covid-19 fake news using biomedical information extraction, Big Data and Cognitive Computing 7 (2023) 46.
[10] J. Dagdelen, A. Dunn, S. Lee, N. Walker, A. S. Rosen, G. Ceder, K. A. Persson, A. Jain, Structured information extraction from scientific text with large language models, Nature Communications 15 (2024) 1418.
[11] S. Luo, J. Yu, Esgnet: A multimodal network model incorporating entity semantic graphs for information extraction from chinese resumes, Information Processing & Management 61 (2024) 103524.
[12] A. Hadji, M.-K. Kholladi, Automatic opinion extraction from football-related social media: A gazetteer and rule-based approach, NCAIA'2023 (2023) 61.
[13] A. Hadji, M.-K. Kholladi, N. Borisova, Enhancing spatial information extraction from arabic text: A hybrid approach with ontology and rule-based, Ingenierie des Systemes d'Information 29 (2024) 1261.
[14] A. Hadji, M. K. Kholladi, Advanced nlp methods for disaster information extraction: Analyzing jape rules, ontologies, and machine learning approaches, in: Proceedings of the 3rd International Conference on Computer Science's Complex System and their Application (CCSA'2024), Computer Science Book Series, Springer Nature, 2024. In press.
[15] F. Gutierrez, D. Dou, S. Fickas, D. Wimalasuriya, H. Zong, A hybrid ontology-based information extraction system, Journal of Information Science 42 (2016) 798–820.
[16] D. Maynard, W. Peters, Y. Li, Metrics for evaluation of ontology-based information extraction., in: EON@ WWW, 2006.