information

MDPI

Article

Automated Social Media Text Clustering Based on Financial Ontologies

Andrea Calvagna *, Emiliano Tramontana and Gabriella Verga

Dipartimento di Matematica e Informatica, University of Catania, 95125 Catania, Italy; tramontana@dmi.unict.it (E.T.); gabriella.verga@unict.it (G.V.)
* Correspondence: andreamario.calvagna@unict.it; Tel.: +39-095-7383008

Abstract: Social media networks provide an aggregation of news and content, allowing users to share and discuss topics of greatest interest to them. Users can enrich the news by providing context and opinions that are useful to other users. Understanding topics of interest sheds light on the collective thinking of a group of individuals and offers important insights for exploring a given field. Among the fields of interest on social media networks, finance stands out. Automatically identifying and organizing the main issues that users discuss can be useful for multiple purposes, e.g., identifying the preferred types of loans could be useful for refining targeted advertising. Our work aims to identify and organize the topics discussed on a social media network that are related to the financial sector. For this, we propose an approach that consists of analyzing posts from Reddit communities oriented to finance. First, posts were gathered and cleaned to remove punctuation, links, and images. Then, textual similarity was computed to match posts with classes from dedicated ontologies designed for the financial sector. Finally, the populated ontology was analyzed to identify clusters of concepts. The results showed that the proposed approach and corresponding tool can summarize topics from a large number of Reddit posts using the identified classes. Over 70% of posts were linked to ontologies when considering both posts and comments, which shows that the automatic support given to posts related to financial concepts had a high degree of success.

Keywords: social media; finance; ontology; data integration; data analysis; classification

Citation: Calvagna, A.; Tramontana, E.; Verga, G. Automated Social Media Text Clustering Based on Financial Ontologies. *Information* **2024**, *15*, 210. https://doi.org/10.3390/info15040210

Academic Editor: Emilio Matricciani

Received: 4 March 2024
Revised: 30 March 2024
Accepted: 6 April 2024
Published: 9 April 2024

1. Introduction

The last decade has witnessed a profound transformation in communication, largely due to the emergence of social networks. Digital platforms have greatly changed human interactions globally and on a systemic level, impacting various aspects of life such as personal, work, social, educational, and political domains. Previously controlled by traditional media and agencies, the flow of information and knowledge is now accessible to the entire population, allowing anyone to publish content or express opinions.

In this context, tools for analyzing social media are incredibly useful. In fact, they provide valuable insights and actionable data that businesses can leverage to improve their efficiency and drive business growth in today's digital landscape. Social media analysis involves the use of various techniques to extract insights, trends, and patterns from social media data [1–4]. Text mining and Natural Language Processing (NLP) techniques are used to analyze textual data from social media posts, comments, and messages. This includes tasks such as sentiment analysis, topic modeling, named entity recognition, and keyword extraction. Sentiment analysis involves determining the sentiment or emotional tone expressed in social media content. This technique classifies text as positive, negative, or neutral, enabling businesses to understand customer sentiment toward their brand, products, or services [5]. Topic modeling techniques, such as Latent Dirichlet Allocation (LDA) [6,7] or Non-Negative Matrix Factorization (NMF) [8,9], are used to identify latent topics or themes in social media discussions. This helps businesses understand the main topics of conversation and emerging trends within their industry or target audience.

The use of an ontology enhances the efficiency of analyzing and retrieving unstructured content in social media data [10,11] and in the medical realm [12,13]. An ontology can enrich the semantics of unstructured text by providing specific conceptual representations of entities, thus improving the accuracy of concept identification [14,15]. Interpreting relations in social media data based on a dataset-specific ontology leads to the effective discovery of inherent relationships between entities [16]. Potential applications include improving the relevance of data retrieved from user profiles on social network websites or developing semantic search engines. However, the effective use of an ontology requires a considerable initial investment due to the development of an accurate domain-specific ontology. This has, in the past, limited the use of ontologies in social media analysis.

In this paper, we present an original approach to implementing automatic topic extraction from social media posts, based on the use of a consolidated and well-known financial domain ontology. This will result in the ability to automatically group social media posts based on the topics they actually cover. Specifically, in this work, we present an approach to identify interest groups within the Reddit community (Reddit APIs allowed us to perform automatic data extraction). Our approach consists of analyzing posts created by various Reddit users within the same subreddit tag, which are linked to financial subjects. Starting from these posts, interest groups are identified through a correlation process, deriving affinity relations with concepts defined in a well-known and widely adopted domain-specific financial ontology.

The automatic linking of posts with ontology classes could be of assistance to financial advisors in various ways: (i) selecting posts by accessing one (or more) ontology concepts; (ii) finding posts that are related to each other by an ontology concept; (iii) representing many posts using a smaller number of ontology concepts; (iv) summarizing and highlighting important aspects of posts connected to an ontology concept; and (v) quickly identifying the most important concepts that users are interested in by looking at the number of posts for each concept. Then, accordingly, financial advisors could read a selection of related posts and possibly propose financial services that may be of interest to users.

This paper is structured as follows. Section 2 offers an overview of the works carried out in the literature. Section 3 presents the key concepts, i.e., the ontologies and external sources used. Section 4 explains our approach. Section 5 presents the results of our analysis. Finally, Section 6 presents the conclusions.

2. Related Works

The state of the art in ontology used for analyzing social media content is characterized by a multidisciplinary approach that integrates ontological knowledge with machine learning, NLP, and domain-specific expertise to extract actionable insights and unlock the full potential of social media data [1–3,17–19]. Ontology integration into the analysis process enhances the efficiency of retrieving unstructured content in social media data [10,11]. Ontologies enrich the semantics of unstructured data by providing specific conceptual representations of entities, thus improving the accuracy of concept identification [14,15]. Interpreting relations in social media data based on a dataset-specific ontology leads to the effective discovery of inherent relationships between entities [16]. Potential applications include improving the relevance of retrieved information on social networking websites based on user profiles and domain understanding or developing semantic search engines. However, the effective use of ontology requires a considerable investment in developing an accurate domain-specific ontology, which, in the past, has limited its use in social media analysis.

Recent works on ontology-based analysis of social media data have focused on several key areas. Alt et al. proposed an approach to increase the efficiency of defining ontologies by automatically extracting knowledge from existing enterprise application systems [20]. Wongthongtham et al. proposed an ontology-based approach focused on extracting the semantics of textual data at the entity level and domain level, using Twitter as a social channel for the concept of proof [14]. An ontological model was presented by Moshkin for

the unification of data profiles of different social networks, avoiding data redundancy and including contextual information in annotations to ontology relations [21].

A semantic approach employing a generic and intelligent framework was proposed by El Kassiri et al. in [22] to respond to different analytical needs applicable to online social network data, leveraging ontologies' inference potential. Moreover, machine learning and deep learning methods have been used for feature engineering in social media data analysis, with a focus on an ontological view of the data to represent knowledge in a more understandable form [23]. The use of ontologies in social media data analysis was emphasized as crucial by El Kassiri [24] to support interoperability and data aggregation in social media. He developed a unified semantic model using standard social ontologies, which can be extended to support future social media.

Other very interesting strategies for text analysis and text clustering include those based on complex networks and graph theory, such as those described in [25,26]. Graph clustering techniques surely merit more attention and could be used in our future work to cluster ideas, also using community detection.

However, key challenges in this area still exist. Some major examples include improving the efficiency of ontology-based sentiment analysis for social data or leveraging entity extraction and concept mappings for accurate concept identification, serving as a dictionary for analyzing unstructured social media content [14].

While existing works have provided valuable insights into various aspects of ontology-based analysis in the context of social media, none of them have directly addressed ontology-based automatic classification of social media posts in the financial domain, despite it being a major topic of interest in social interactions today for many social network users.

3. Background

3.1. The Financial Industry Business Ontology

The Financial Industry Business Ontology (FIBO) is the global standard ontology for efficient and unambiguous financial services [27]. FIBO (https://spec.edmcouncil.org/fibo last accessed 28 February 2024) contains semantic links between financial concepts, describing their meanings, and is intended for practical use in the real world. In particular, FIBO describes basic concepts used in the financial world, such as legal entities and financial processes [28]. In practice, FIBO is not a single ontology but rather a set of ontologies divided into modules and submodules. The modules include the following areas:

1. Foundations (https://spec.edmcouncil.org/fibo/ontology/FND/MetadataFND/FNDDomain last accessed 28 February 2024) contains ontologies belonging to the Foundations (FND) domain, which define general-purpose concepts required to support other FIBO domains. These include concepts and relationships about people, organizations, places, and, most importantly, contracts essential to domains such as Business Entities (BE), Financial Business and Commerce (FBC), Indices and Indicators (IND), and Securities (SEC). It organizes the knowledge of 66 ontologies.
2. Business Process (https://spec.edmcouncil.org/fibo/ontology/BP/MetadataBP/BPDomain last accessed 28 February 2024) contains ontologies belonging to the Business Process (BP) domain, which define financial process flows, such as securities issuance and transaction workflows. In the case of securities issuance process models, these are provided to represent reference data concepts dependent on the process by which a security was issued. Transaction process semantics provide the basis for the temporal dimension of securities and derivatives transactions. It organizes the knowledge of 11 ontologies.
3. Indices and Indicators (https://spec.edmcouncil.org/fibo/ontology/IND/MetadataIND/INDDomain last accessed 28 February 2024) contains ontologies belonging to the FIBO Indices and Indicators (IND) domain. This domain covers market indices and reference rates, including economic indicators, foreign exchange, interest rates, and other benchmarks. The ontologies cover quoted interest rates, economic measures such as employment rates, and quoted indices required to support baskets of securities,

including specific kinds of securities in share indices or bond indices, as well as credit indices. It is constituted by 20 ontologies.

4. Derivatives (https://spec.edmcouncil.org/fibo/ontology/DER/MetadataDER/DERDomain last accessed 28 February 2024) contain ontologies belonging to the Derivatives (DER) domain. This domain covers many of the concepts common to derivative instruments, including but not limited to options, futures, forwards, swaps, and a wide range of other derivatives. It includes the knowledge of 24 ontologies.

3.2. The R/Wallstreetbets Community

Reddit is a popular medium for exchanging ideas and communicating about various fields. Due to the large number of topics, Reddit is made up of thousands of smaller communities, referred to as subreddits. Several communities related to the financial field have been created on Reddit. As of February 2024, the largest subreddits dedicated to investing or trading were r/wallstreetbets (13.5 million subscribers), r/stocks (5.1 million), and r/investing (2.1 million) [29]. The r/wallstreetbets (www.reddit.com/r/wallstreetbets/ last accessed 28 February 2024) community is widely used by different types of users because, although purely financial, the community overcomes the barriers caused by financial jargon, i.e., users familiar with the jargon explain ideas to the masses in simpler terms. In this paper, we processed posts extracted from the *r/wallstreetbets* community since it is the largest currently available source of Reddit post data. However, the proposed approach is general and can be applied to any set of posts, any community, or any group of communities. There are no particular selection criteria for the posts that can be processed, other than removing any personal information and skipping posts with images as the main content. The whole set of posts used to test this work is still available on Reddit as of today and has been made publicly available as a shared .cvs file (https://github.com/amcalvagna/REDDIT-Source-Data, accessed on 30 March 2024).

4. Proposed Approach

In our proposed approach, the text written in the posts is analyzed to determine their possible associations with the concepts expressed in FIBO, a set of ontologies dedicated to the financial field. The aim is to automatically identify and summarize the recurrent financial concepts that are of interest to the users. Figure 1 gives an overview of the approach. The following processing steps are performed to organize text from posts. The association between posts and concepts is the result of step 3, whereas the summarization of recurrent concepts used in posts is the result of step 4.

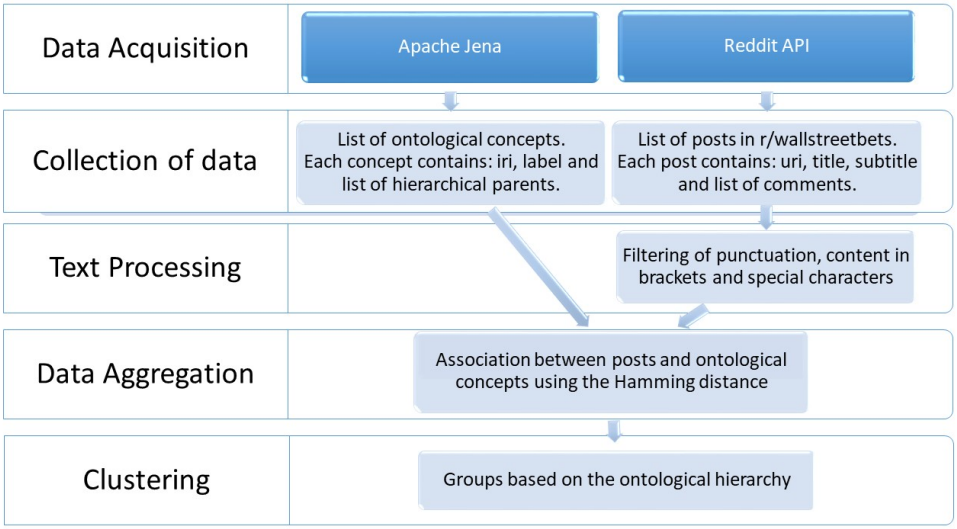

Figure 1. Overview of the proposed approach.

The steps performed were as follows:

1. Using the APIs provided by the Reddit platform, the text in the posts of the selected subreddits was gathered.
2. The ontologies in FIBO were parsed to extract the names of the financial concepts represented in them as classes.
3. The text from the Reddit posts was cleaned (essentially to remove punctuation), and then the degree of connection between each post and a financial concept present in the FIBO ontologies was computed.
4. Posts that were connected to some FIBO classes were further analyzed to identify potential clusters, which could be determined by concepts represented in the FIBO ontologies appearing together frequently across the posts.

These steps are described in the following section.

4.1. Reddit Data Extraction

Reddit offers the possibility to read posts for free and for research purposes. The Python community provided the PRAW package (https://pypi.org/project/praw/ last accessed 28 February 2024), an acronym for Python Reddit API Wrapper, that allows simple access to the Reddit API. To use the package, the user needs credentials for a script-type OAuth application, which are provided after registering on the Reddit platform and creating a project (https://www.reddit.com/dev/api/ last accessed 28 February 2024). We implemented a getAPI(x,y) method that returns a list of posts, where x is a subreddit and y is the number of desired posts. This method calls the given openConnection() method, which uses four parameters necessary for connection with Reddit and provided by the Reddit system when registering the project: username, password, client_id, and client_secret. A relevant snippet of code for the getAPI() method is presented in Listing 1.

Listing 1. Pseudo code for the getAPI algorithm.

```
def getAPI(subreddit_name, num):
    reddit = openConnection()
    subreddit = reddit.subreddit(subreddit_name)
    posts = subreddit.new(limit=number)
    return posts
```

From the list of posts returned, for each post, the following parts were extracted:

- URI: unique identifier of the post.
- Title: identifies the main topic of the post in a brief and concise manner.
- Subtitle: provides a more detailed descriptive text of the topic.
- Comments: discussions written by other users.

All posts gathered were from the r/wallstreetbets community.

4.2. FIBO Parsing and Data Extraction

In general, ontologies consist of connected concepts that form a tree, where the root represents the most generic concept and the leaves represent more specific concepts. Therefore, for each ontology, the list of classes (concepts) was obtained, and for each class, the following parts were collected:

- URI: unique identifier of the concept.
- Label: name identifying the concept.
- List of superclasses: list of more generic concept URIs.

Apache Jena is a free and open-source Java framework for building semantic web and linked data applications, and it has proven valuable for developing systems working with ontologies [30,31]. Apache Jena was used to read and navigate the ontologies and

retrieve all the class labels in them. The implemented getAllClassesByURI(String uri) method analyzed an ontology defined by the URI parameter. The Apache Jena framework downloaded the ontology and identified the list of classes in it. For each class (concept), the name of the class, the definition property, and the closest connected classes were obtained by exploring the ontology tree.

A snippet of code for the getAllClassesByURI() method is presented in Listing 2.

Listing 2. Pseudo code for the getAllClassesByURI method.

```
procedure getAllClassesByURI(String uri){
    OntModel model = ModelFactory.createOntologyModel(OntModelSpec.OWL_DL_MEM);
    model.read(uri);
    for each ontology class in the model do {
      filter out all the class terms in its description note and
      find and collect the classes one level up over them in the ontology
      add these to the set of collected class labels
    }
   return the whole set of class labels
}
```

All the ontologies of the macro-areas listed in Section 3.1 were extracted: FND, BP, and DER. Each ontology contained a list of IRIs associated with it. For example, within FND, there is an ontology designed for Contracts available from the IRI FND/Agreements/Contracts.rdf and an ontology designed for Jurisdiction available from the IRI FND/Law/Jurisdiction.rdf (each IRI is preceded by a path, which is the same for all the above, i.e., https://spec.edmcouncil.org/fibo/ontology/ last accessed 28 February 2024).

4.3. Connecting Text from Posts to FIBO Classes

Posts extracted from Reddit were analyzed to determine the degree of affinity with ontological concepts. First, the text gathered from the posts (comprising title, subtitle, and comments) was cleaned by deleting punctuation and some textual parts deemed useless, such as round and square brackets and the text within them. Punctuation is unnecessary for matching words, hence it was removed. The text found within brackets, which was removed, often contained web links, which are not useful for matching words. Second, each concept expressed in the ontologies, that is, the label of a class, was compared with the text of the posts. For this comparison, the number of words obtained from the sentences in posts was adjusted to match the number of words that constituted a label in the ontologies, while varying the starting point in the sentences to allow for all possible comparisons with labels.

The metric used to compare strings was the Hamming distance [32]. The Hamming distance for two strings of equal length was computed as the number of positions in which the corresponding symbols were different, given by the number of substitutions needed to convert one string to the other. For example, a pair of words with a Hamming distance equal to 0 or 1 were considered similar. This means they could have had a similar meaning even though small variations occurred, such as for love and lover. If the distance was equal to 0, the words or group of words were identical; if the distance was equal to 1, they differed by one character and could be combined, e.g., price and its plural prices. Pseudo code for this string distance computation is shown in Listing 3.

Listing 3. Pseudo code for the Hamming distance computation method.

```
def distance_hamming(chain1, chain2):
    if abs(len(chain2) - len(chain1)) <= 1:
        return sum(c1 != c2 for c1, c2 in zip(chain1, chain2))
```

4.4. Clustering Concepts and Posts

For each class of the FIBO ontologies associated with some posts, the closest hierarchy of the class was identified by carrying out a bottom-up navigation of the tree, by determining the father, called a superclass; grandfather (father of the superclass); and great-grandfather (father of the grandfather). This hierarchy was composed of the three levels of the tree, with each level connecting one class (concept) with another among the closest classes (another concept). Using this hierarchy, concept clusters were created, where all related classes had a common ancestor class.

The searchAllChildren() method, whose pseudo code is shown in Listing 4, was implemented to analyze all the classes that were associated with a post and shared a common ancestor. Parameter list_classes is a dictionary that was built to contain the classes that belonged to the same ontology and had associations with posts.

Listing 4. Pseudo code for the searchAllChildren method.

```
def searchAllChildren(uri, list_classes):
    list_children = []
    for element in list_classes:
        if uri in element:
            list_children.append(element)
    return list_children
```

The ontological classes within a cluster represented well-described concepts in the ontology. Each class, by construction, was associated with posts. Thus, by analyzing the concepts in a cluster, it was possible to determine the areas of interest of the users who submitted the posts. For example, as shown in more detail in the following section, one cluster identified people and included classes such as organization member, issuer, owner, partner in a partnership, shareholder, etc. Another cluster included classes such as ownership, control, possession, affiliation, beneficiary, and other similar relationships.

4.5. Note on the Accuracy and Methodology

The proposed fully automated approach for organizing posts on Reddit is novel. No previous dataset exists that provides ground truth for organizing a dataset into ontology concepts. Curating a dataset that provides ground truth could be a contribution in itself. For this, one would need deep knowledge of the ontology to tell concepts apart and an understanding of the posts. Of course, it would take a considerable amount of time to manually analyze thousands of posts. Importantly, it should be noted that we do not use semantic groups to create a hard partitioning of the posts, that is, we do not aim to identify neatly separate clusters of posts. Based on the proposed process, it is pointless to try to assess categorization accuracy as it does not involve a classical classification task. That is, it is not intended to determine post-data self-similarity and divide them into distinct categories based on complex criteria. Instead, the identification and population of semantic groups, along with their association with related posts, are based on string pattern matching. We determine semantic categories by matching post contents exactly with ontology entries. Then, we process the posts to find their connections with each of these semantic groups, again using string matching. As a consequence, it is common for a post to be associated with not just one but multiple semantic groups. This is not a processing step that can introduce imprecision, since it yields only deterministic boolean outcomes: either a term is used in a post or is not, and thus a relation exists or does not. Moreover, if several possible connections of posts to classes could be considered correct, given the similarity of concepts expressed in some classes and the ambiguity/vagueness of some posts, we have addressed the possible ambiguity of posts by searching the context of words when such words have several meanings, before attempting a connection to an ontology class. The semantic groups themselves thus become a tool for indexing, navigating, searching, or summarizing the considered set of posts, extending beyond their original community affiliations.

These ontological groups are not designed to partition the post into disjoint clusters but to organize them in possibly partially overlapping clusters, each with a well-defined semantic tagging coherent with the current social media content.

5. Results and Discussion

In this section, we present the results of our analysis of posts extracted from the Reddit platform and connected to FIBO ontologies. Table 1 shows a small sample of posts with titles, subtitles, and the associated labels from the ontologies. Among all the posts gathered from Reddit (over 4500) and the r/wallstreetbets community, 2700 distinct posts were retained since they contained text in the titles and subtitles (posts with only a title and a picture were not considered). The FIBO ontologies used were those found in the FND, BP, and DER macro-areas.

First, we formed a collection of posts, each containing a title and a subtitle. The connections of these posts to the labels from the ontologies were searched, and then clusters for classes with connections were created. This first experiment was called Test 1. Second, we formed a collection of posts, each containing a title, a subtitle, and some comments. Then, the connections of these posts to the labels from the ontologies were determined, and clusters were created (similar to before). This second experiment was called Test 2.

Table 1. Each row shows the title of a post, its subtitle, and the list of ontology labels found for it.

Post Title	Post Subtitle	Labels
We've seen your picks for 2024 stocks, what about which ones are going to be flat?	In general, making money off movement with calls and puts is good, but I also really appreciate bets on **prices** ending up level. Personally I think SPX could end flat in 2024, or marginal movement either way, so I plan on trading a SPY butterfly for EOY 2024. Products that I could see ending the year flat in 2024 are: metals like gold and silver, PYPL (sorry), ETSY, and C. I'm thinking 2024 will make an "M" shape. Optimism on rate cuts, rate cuts happen but it's buy the rumor sell the fact, then more optimism, followed by a reality check. Credentials: none. Love regards. Positions: long term hold on SPY shares like everyone else, nothing else in anything else mentioned, but I **trade** metals frequently and trade active stocks.	price, trader, trade
100 Years of Gold vs. Stocks	Companies are a **representation** of human ingenuity and productivity that you can own. Companies can evolve from a few people working out of a garage to international businesses with manufacturing. intellectual property. and dare I say **precious metal** reserves .. Generally human ingenuity is better to invest in than a rock. Even if that rock is way better than holding cash. .. Imagine trying to flee a failing **country** with $1 million of gold vs. $1 million of crypto. Imagine if we solve nuclear reactions to create gold. Imagine if we get good at asteroid mining.	country, precious metal, representation
Unity3d laid off 1/3 of its employees realizing it invalidated all its contracts	Users don't even have to pay anymore. Bankruptcy **future**? I **broke** the news on /r/wallstreetbets about Unity floundering as soon as they mistepped. and as you can see many disagreed with me ... or just dumb automatic **scripts** taking in losses ... ZombieGorrila an official Unity **employee** brought up religion as a reason to ban me when I never mentioned anything religious. Unity3d is a ship of Thesus and could be salvaged with proper management. but I don t **bank** on it... Unity looks like a good sell. an astute and knowledgeable short **seller** might make money if they make sure to: A: let the dust settle if the efficiency layoffs make sure it didn t stop the bleeding and B: Immediately buy back IF A BIG PLAYER SWOOPS IN... He just made terrible mistakes in leadership. culture. and maintaining a **product** people loved and still do... give Unity the **right** to change aspects of the **contract**... BUT they would never invalidate the original **contract** over this.. source on the breach of **contract** thing?. Unless you are a lawyer that literally writes terms of services **contracts** ... Can anyone else confirm whether the breach of **contract** is actually making them lose money or not. This guy is schizo and all his unity puts is going to zero.. What were the terms in the **contracts** you think they are breaking?	script, employee, contract, good, seller, product, right, bank, future, broke.

Using the collection formed in Test 1, labels from ontologies were associated with approximately 1200 posts. In this case, the cluster consisting of the highest number of labels contained 20 labels, whereas the cluster connecting the highest number of posts contained 432 posts. From the collection formed in Test 2, labels were associated with approximately 1700 posts. In this case, the cluster consisting of the highest number of labels contained 25 labels, whereas the cluster connected to the highest number of posts contained

1508 posts. Table 2 shows a small sample of clusters, indicating the representative label for each cluster, a description of the label, the number of labels within the cluster, and the number of connected posts.

Some words were used in the posts with a different meaning than the ontological labels. For example, good, future, and holding have very different meanings in the two sources (posts and ontological labels). Therefore, we listed common words that have several meanings, and we filtered out these words when looking for a connection between posts and ontological labels to avoid false positives.

Figure 2 shows an example of a cluster identified in the ontology FND/TransactionsExt/MarketTransactions.rdf. The cluster is illustrated using a tree structure. The root node of the tree (the box on the left) shows the URI label of a class (in capital letters within the box) along with its definition (words in the box below the label). This root node is inherited by other URIs (in the center). The yellow node is a node that is inherited by other nodes (on the right). Each node shown is connected to posts obtained from Reddit. For each node, the number of posts connected to each ontology class is shown within round brackets beside the label.

Table 2. Each row represents a cluster comprising a cluster identifier, a significant class (label), the description of the significant class, the number of labels within the cluster, and the number of posts connected to the cluster.

Cluster	First Label	Description	Labels	Posts
1	party-in-role	Relative concept that ties a person or organization to a specific role they stand in.	20	166
2	agent-in-role	Relative concept that ties an agent to a part they play in a given situational context.	16	151
3	thing-in-role	Relative concept that ties something to a part it plays in a given situational context.	14	217
4	situation	Setting, state of being, or relationship that is relatively stable for some period of time.	8	227
5	independent party	Any person or organization.	7	30

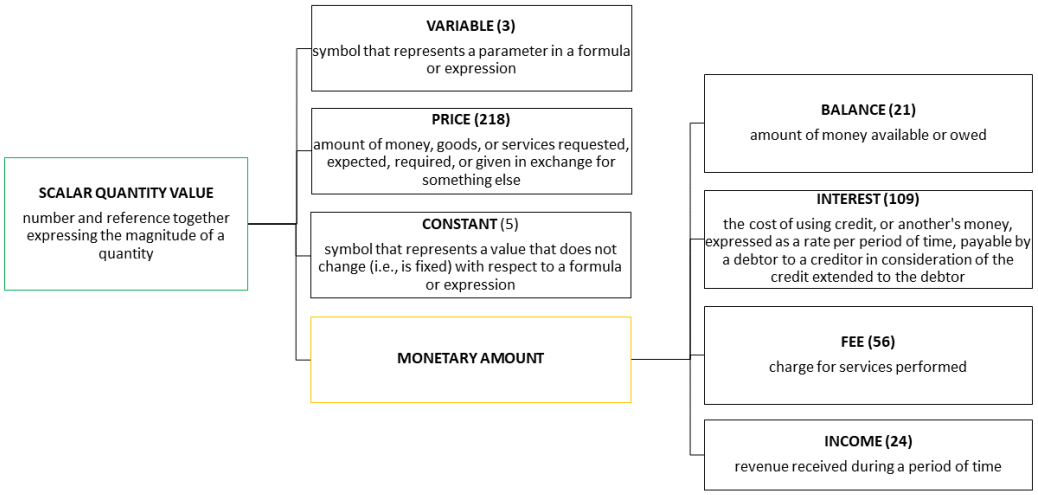

Figure 2. A sample cluster of classes showing the main class on the left, its URI, and the inherited classes with the number of connected posts.

Table 3 shows some results obtained from both tests. The two results highlighted in bold in Table 3 refer to the Monetary Amount label, which is both a label and an identified cluster. Although the cluster identified in the two tests was the same, as defined by its constituting classes, the number of connected posts differed. There were 222 posts connected in the first test, and 897 posts connected in the second test. This was due to the fact that Test 2 included a higher number of posts, including those with comments.

Table 3. Some clusters resulting from both tests, the number of connected posts, and the constituting labels.

Test	Main Label	Posts	Constituting Labels
1	Scalar Quantity Value	1058	price, fee, balance, interest, constant, income, variable
1	Document	390	notice, record, report, catalog, publication, license, certificate, passport, appraisal
1	Actor	297	investor, owner, receiver, executive
1	Legal Construct	127	right, claim, regulation, duty
1	**Monetary Amount**	**222**	**fee, interest, income, balance**
2	Functional Entity	519	trader, government, dealer, merchant, underwriter
2	Expression	813	total, median, mean, percentage, ratio, variance
2	Obligor	13	payer, borrower, debtor
2	Financial Service Provider	252	bank, underwriter
2	**Monetary Amount**	**897**	**fee, interest, income, balance**

By analyzing the definitions and concepts that define a cluster, it is possible to better understand and study the fields of interest of users. Each cluster provides the labels that facilitate connections with posts. Over the long term, by observing the change in the number of posts connected to a cluster, it is possible to determine whether the interests of groups of people remain consistent or have shifted elsewhere. Moreover, by analyzing the resulting clusters, it is possible to discover related interests (common to a cluster) and create recommendation systems.

Figure 3 shows two examples of clusters presented as trees of classes. On the left, the root URI has the label scalar quantity value, and the cluster was identified within the ontology FND/TransactionsExt/MarketTransactions.rdf. The hierarchy shown is the same as that of the cluster shown in Figure 2 (last accessed 30 March 2024). On the right, the root URI has the label commitment, and the cluster contains debt, warranty, and representation from the same ontology.

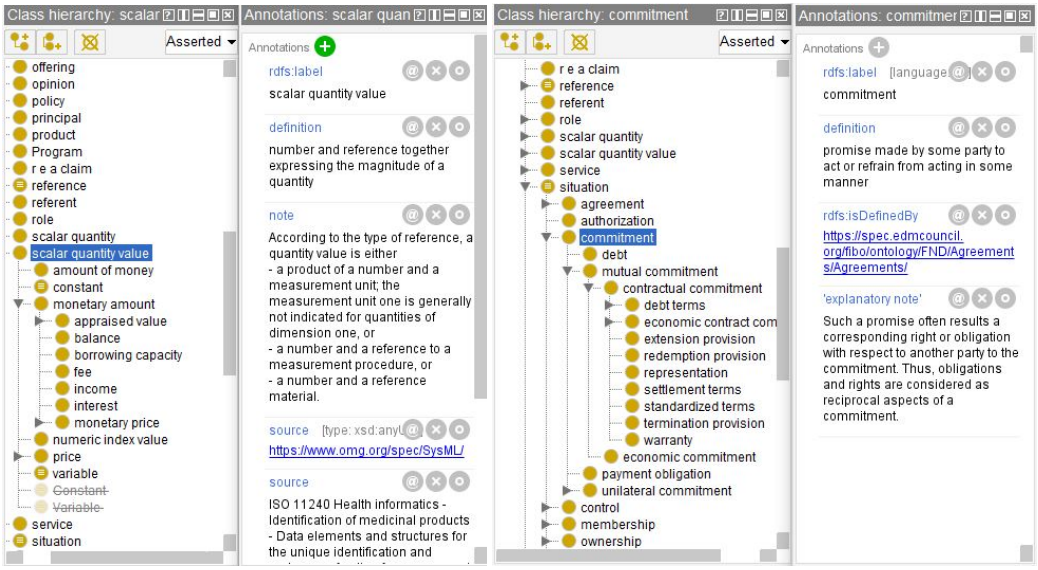

Figure 3. Two examples of identified clusters: one for scalar quantity value, on the left, and another for commitment, on the right.

Some clusters were detected based on information present in multiple ontologies. For example, there was a cluster with the parent URI documents and the following terms identified in the posts: notice, record, report, catalog, publication, license, certificate, passport, and appraisal. This cluster resulted from the combination of labels present in the three ontologies.

Overall Analysis of the Processed Posts

In this section, we analyze the additional processed data to highlight how the size and composition of the semantic groups and the number of connected posts are related

and can provide insights into interesting aspects of the active discussions about the considered domain: What are the most active topics? Which group of discussion topics exhibits the most accurate, technically verbose, and in-depth lexicon for dealing with the related subject?

We provide a series of graphic charts representing the distribution of data resulting from Test 1 performed on Reddit post contents. We start by showing in Figure 4 an overall graphic chart representing the size of each identified semantic cluster (groups) in terms of closely connected ontology classes (labels) and their respective number of connected posts, sorted by descending post count.

It should be noted that the size of the post count follows a logarithmic distribution with respect to the ontology groups, which means that there is a very small set of concept groups consistently receiving significantly more attention than the others. Also, we can see that a large majority of the groups receive very little attention from the community.

We assume that semantic groups with a poor vocabulary and very few posts are symptomatic of trivial, superficial, or inherently irrelevant discussion topics. It also might erroneously be concluded that this post distribution is not related to the size of their clusters, but this is a consequence of these figures being very small compared to the post counts.

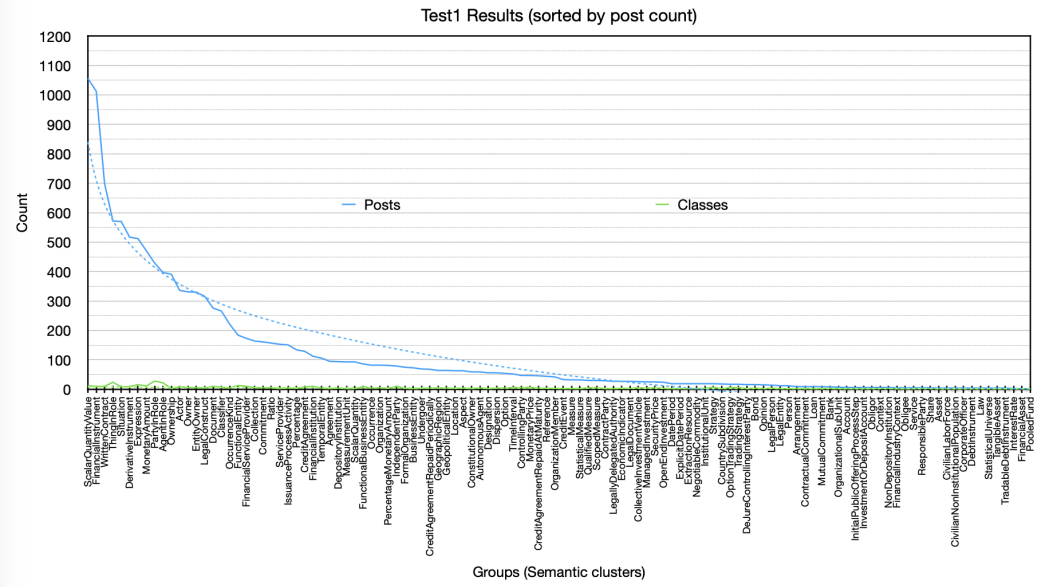

Figure 4. Graphic chart representing the size of each identified group in terms of closely connected ontology classes and their respective number of connected posts. The dashed line represents the trend computed as the logarithmic interpolation of the data series.

In fact, it is evident that this is not the case from looking at Figure 5, which redraws the same quantities on a logarithmic scale while also reordering the data series with respect to another metric: the normalized cluster size. This metric represents the ratio of connected posts vs. the size (in terms of the number of classes) of the respective cluster, providing a measure of the popularity of the post irrespective of the size (or verbosity) of their associated group. The sizes of the groups in this re-scaled view clearly exhibit a logarithmic decreasing trend, similar to that of their related post counts. Despite the variability in their absolute data values, this logarithmically scaled chart clearly shows a direct proportionality relationship between the average sizes of the ontology groups and their respective number of posts. This is not surprising at all; the proportionality relationship between the two data series is actually already defined by the normalized cluster size by construction, and this parameter exhibits a smooth linear trend. This could make it a good candidate as a parameter to set thresholds for the selection of the best groups of posts to process for further specific applications.

In Figure 6, we present the whole set of ontology labels actually used, along with the number of connected posts found and the corresponding body of text for all those posts. This shows how a small fraction of semantic labels receive hundreds of times more attention in posts than all the others, resulting in a proportional but much larger amount of social media text content, in terms of words to be read. This is a clear indication of the significance and interest to users. As a consequence, it is of great convenience to be able to identify the ontological concepts related to that large body of text and to gain a quick understanding of the type of content it represents by simply examining a short list of connected terms instead of having to go through hundreds of thousands of words to catch the details.

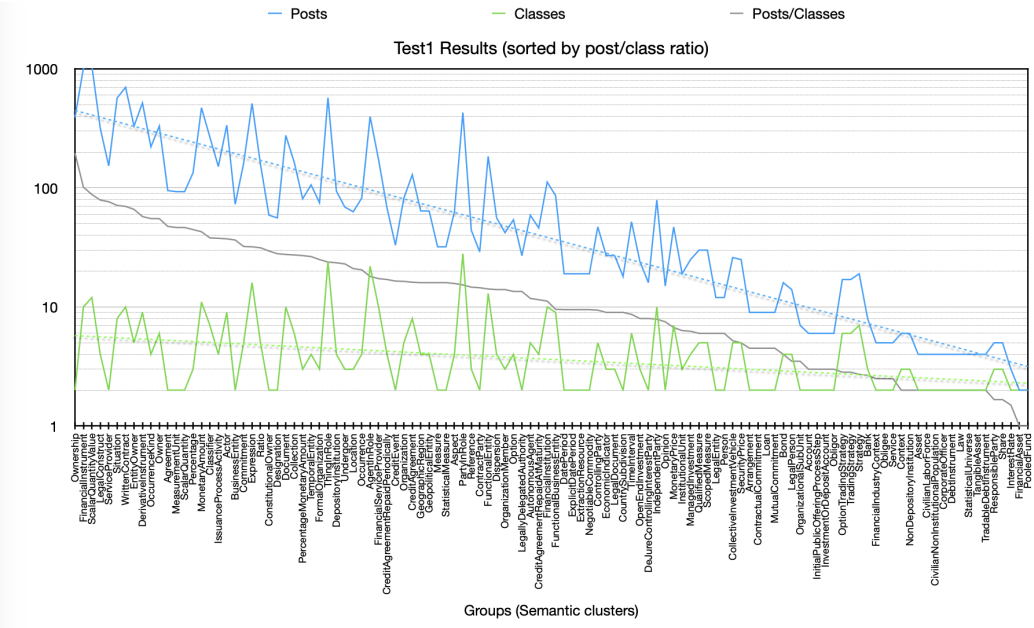

Figure 5. The normalized cluster size. This chart shows a logarithmic scale representation of data collected in Test 1 for the identified ontology groups. The x-axis is sorted by descending post/class ratio, whereas the y-axis shows the post count and class count for each group. The trend lines for the post and class counts are indicated by dashed lines, computed as the exponential interpolation of the respective data series. The normalized cluster size is shown as a gray line and computed as the ratio between the post count of an ontology group and its cluster size, representing the number of closely related technical terms it represents. The chart demonstrates that this proportionality actually has a logarithmic distribution, with groups up to a hundred times more significant in this respect than others.

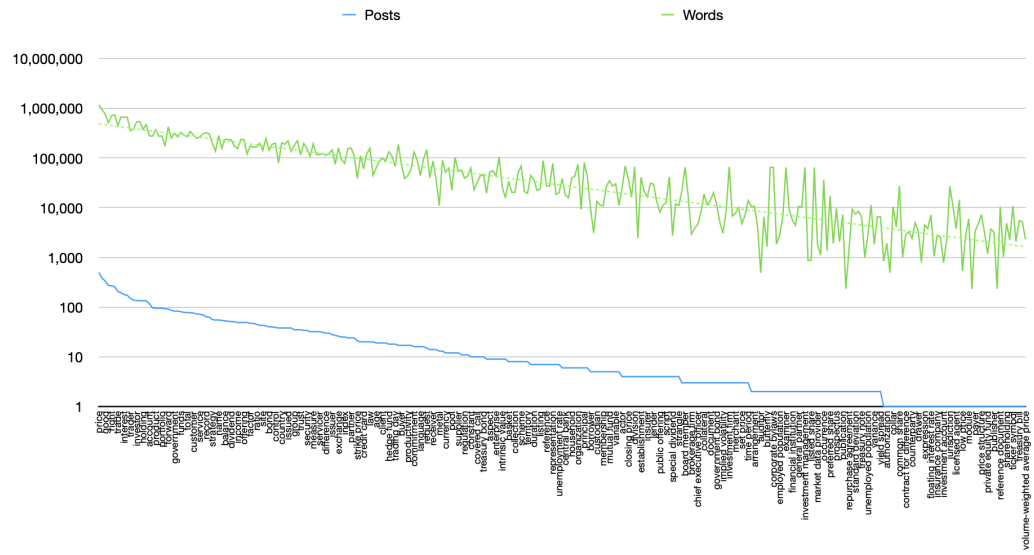

Figure 6. Posts per single class label and the corresponding size in words of its overall body of text. The trend line for the size in words count is also shown as a dashed line, computed as the exponential interpolation of the corresponding data series

6. Conclusions

This study explored the automatic aggregation of discussion topics on social media platforms, particularly focusing on those related to financial matters. While there exist several works on ontology-based analysis of unstructured text, none of them directly address the automatic classification of social media posts in the financial domain.

We presented an original approach to achieving this goal and also implemented and tested it on actual data extracted from a large volume of Reddit social media posts.

Reading and organizing thousands of user posts would be infeasible within a reasonable amount of time. The proposed tool automatically organizes the posts by connecting them to ontology classes, with each class providing the number of connected posts and the actual posts themselves.

We showed that the devised approach is able to automatically process large sets of unstructured social media text, effectively identifying key discussion topics and clustering relevant and semantically related concepts. In fact, by using financial concepts from a specialized ontology, we could highlight meaningful keywords and financial concepts that may have predictive power or practical implications. The ontological connections assisted in identifying interested user groups while covering a significant portion of the posts.

The proposed implemented approach provided valuable insights for understanding collective thinking within the financial sector, including aspects well beyond the discovery of user special interests or the support of recommendation systems. Indeed, there are limitations to analyzing and organizing social media data. As pointed out by Gore et al. [33], social media messages are posted by an unrepresentative portion of the population and thus contain demographic and geographic biases that can contribute to distorted views. While it is reasonable to assume these biases also exist in financial posts, we must consider that our analysis is not aimed at extrapolating general statements representative of the whole population. Our analysis aims to profile specific post contents, allowing for a more convenient indexing of their actual contents.

Financial advisors could gain insights from the way the posts are organized and the summaries provided by each class (and group of classes). This organization provides a way to skim posts. Moreover, the number of posts for each class could be used as a priority indicator for advisors when selecting concepts and posts to read. The organized posts could indicate the importance of certain concepts to users, allowing financial advisors to determine the main user interests.

It is important to note that the represented data series is a snapshot of the social media data collected for one run of our tool, whereas these data continuously change over time due to the natural evolution of the local and global financial markets. However, while the groups and actual data figures will change dynamically, the type of data distribution will always highlight a very small subset as more active/interesting than others at each point in time, which our approach will be able to identify.

Additionally, by integrating knowledge from a domain ontology, the topic clustering automatically reflects and defines a network of semantic relations among the contents of the actual post, free from biases imposed by any pre-compiled list of flat options. This is of paramount importance when dealing with datasets whose specific content profile is subject to continuous and dynamic changes over time. As a direction for future development, it is, in fact, of primary interest to focus on the dynamics of how financial topics gain or lose social media traction in the long term, that is, to track the popularity of individual ontology clusters and classes over time.

Author Contributions: Conceptualization, A.C. and E.T.; methodology, A.C. and E.T.; software, G.V.; validation, G.V. and E.T.; writing—original draft preparation, A.C. and G.V.; writing—review and editing, A.C., G.V., and E.T.; supervision, A.C. and E.T. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The whole set of posts used to test this work is publicly available as a .cvs file at https://github.com/amcalvagna/REDDIT-Source-Data.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Kaplan, A.M.; Haenlein, M. Users of the world, unite! The challenges and opportunities of Social Media. *Bus. Horiz.* **2010**, *53*, 59–68. [CrossRef]
2. Kosinski, M.; Stillwell, D.; Graepel, T. Private traits and attributes are predictable from digital records of human behavior. *Proc. Natl. Acad. Sci. USA* **2013**, *110*, 5802–5805. [CrossRef] [PubMed]
3. Mislove, A.; Marcon, M.; Gummadi, K.P.; Druschel, P.; Bhattacharjee, B. Measurement and analysis of online social networks. In Proceedings of the ACM SIGCOMM Internet Measurement Conference, IMC, San Diego, CA, USA, 24–26 October 2007; pp. 29–42. [CrossRef]
4. Jagrič, T.; Herman, A. AI Model for Industry Classification Based on Website Data. *Information* **2024**, *15*, 89. [CrossRef]
5. Calvagna, A.; Tramontana, E.; Verga, G. Revealing People's Sentiment in Natural Italian Language Sentences. *Computers* **2023**, *12*, 241. [CrossRef]
6. Hsu, W.S.; Poupart, P. Online Bayesian Moment Matching for Topic Modeling with Unknown Number of Topics. In *Advances in Neural Information Processing Systems*; Lee, D., Sugiyama, M., Luxburg, U., Guyon, I., Garnett, R., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2016; Volume 29.
7. Tang, J.; Meng, Z.; Nguyen, X.L.; Mel, Q.; Zhang, M. Understanding the limiting factors of topic modeling via posterior contraction analysis. In Proceedings of the International Conference on Machine Learning (ICML), Beijing, China, 21–26 June 2014; Volume 1, pp. 337–345.
8. Fu, X.; Huang, K.; Sidiropoulos, N.D.; Ma, W.K. Nonnegative Matrix Factorization for Signal and Data Analytics: Identifiability, Algorithms, and Applications. *IEEE Signal Process. Mag.* **2019**, *36*, 59–80. [CrossRef]
9. Gan, J.; Liu, T.; Li, L.; Zhang, J. Non-negative Matrix Factorization: A Survey. *Comput. J.* **2021**, *64*, 1080–1092. [CrossRef]
10. Lubis, A.R.; Nasution, M.K.M.; Sitompul, O.S.; Zamzami, E.M. Obtaining Value from the Constraints in Finding User Habitual Words. In Proceedings of the International Conference on Advancement in Data Science, E-Learning and Information Systems (ICADEIS), Lombok, Indonesia, 20–21 October 2020. [CrossRef]
11. Millham, R.; Thakur, S. The Human Element of Big Data: Issues, Analytics, and Performance. In *Social Media and Big Data*; Chapman and Hall/CRC: Boca Raton, FL, USA, 2016; pp. 179–193. [CrossRef]
12. Lazarova, S.; Petrova-Antonova, D.; Kunchev, T. Ontology-Driven Knowledge Sharing in Alzheimer's Disease Research. *Information* **2023**, *14*, 188. [CrossRef]
13. Tripodi, I.J.; Schmidt, L.; Howard, B.E.; Mav, D.; Shah, R. A Tissue-Specific and Toxicology-Focused Knowledge Graph. *Information* **2023**, *14*, 91. [CrossRef]

14. Wongthongtham, P.; Salih, B.A. Ontology-based approach for identifying the credibility domain in social Big Data. *J. Organ. Comput. Electron. Commer.* **2018**, *28*, 354–377. [CrossRef]
15. Calcagno, S.; Calvagna, A.; Tramontana, E.; Verga, G. Merging Ontologies and Data from Electronic Health Records. *Future Internet* **2024**, *16*, 62. [CrossRef]
16. Braun, R.; Esswein, W. OntOSN—An integrated ontology for the business-driven analysis of online social networks. In *Knowledge, Information and Creativity Support Systems*; Kunifuji, S., Papadopoulos, G.A., Skulimowski, A.M., Kacprzyk, J., Eds.; Springer International Publishing: Berlin/Heidelberg, Germany, 2016; Volume 416, pp. 317–334. [CrossRef]
17. Gruber, T.R. A translation approach to portable ontology specifications. *Knowl. Acquis.* **1993**, *5*, 199–220. [CrossRef]
18. Gruber, T.R. Toward principles for the design of ontologies used for knowledge sharing. *Int. J. Hum. Comput. Stud.* **1995**, *43*, 907–928. [CrossRef]
19. Weng, J.; Lim, E.P.; Jiang, J.; He, Q. TwitterRank: Finding topic-sensitive influential twitterers. In Proceedings of the ACM International Conference on Web Search and Data Mining (WSDM), New York, NY, USA, 3–6 February 2010; pp. 261–270. [CrossRef]
20. Alt, R.; Wittwer, M. Towards an ontology-based approach for social media analysis. In Proceedings of the European Conference on Information Systems (ECIS), Tel Aviv, Israel, 9–11 June 2014.
21. Moshkin, V. Unification of Social Media Data When Building a Graph Knowledge Base. In Proceedings of the International Multi-Conference on Industrial Engineering and Modern Technologies, FarEastCon, Vladivostok, Russia, 6–9 October 2020. [CrossRef]
22. El Kassiri, A.; Belouadha, F.Z. A semantic approach towards online social networks multi-aspects analysis. In *Innovations in Bio-Inspired Computing and Applications (IBICA 2017)*; Advances in Intelligent Systems and Computing; Abraham, A., Haqiq, A., Muda, A.K., Gandhi, N., Eds.; Springer: Cham, Switzerland, 2018; Volume 735, pp. 157–168. [CrossRef]
23. Jain, S.; Dalal, S.; Dave, M. An Ontology for Social Media Data Analysis. In *Semantic Intelligence*; Lecture Notes in Electrical Engineering; Jain, S., Groppe, S., Bhargava, B.K., Eds.; Springer: Singapore, 2023; Volume 964, pp. 77–87. [CrossRef]
24. El Kassiri, A.; Belouadha, F.Z. Towards a unified semantic model for online social networks to ensure interoperability and aggregation for analysis. In *Graph Theoretic Approaches for Analyzing Large-Scale Social Networks*; IGI Global: Hershey, PA, USA, 2018; pp. 267–292. [CrossRef]
25. Corrêa, E.A.; Amancio, D.R. Word sense induction using word embeddings and community detection in complex networks. *Phys. A Stat. Mech. Its Appl.* **2019**, *523*, 180–190. [CrossRef]
26. Stella, M.; Zaytseva, A. Forma mentis networks map how nursing and engineering students enhance their mindsets about innovation and health during professional growth. *PeerJ Comput. Sci.* **2020**, *6*, e255. [CrossRef] [PubMed]
27. Bennett, M. The financial industry business ontology: Best practice for big data. *J. Bank. Regul.* **2013**, *14*, 255–268. [CrossRef]
28. Petrova, G.; Tuzovsky, A.; Aksenova, N.V. Application of the Financial Industry Business Ontology (FIBO) for development of a financial organization ontology. *J. Phys. Conf. Ser.* **2017**, *803*, 012116. [CrossRef]
29. Reichenbach, F.; Walther, M. Financial recommendations on Reddit, stock returns and cumulative prospect theory. *Digit. Financ.* **2023**, *5*, 421–448 [CrossRef] [PubMed]
30. Siemer, S. *Exploring the Apache Jena Framework*; George August University: Göttingen, Germany, 2019.
31. Tramontana, E.; Verga, G. Ontology Enrichment with Text Extracted from Wikipedia. In Proceedings of the 5th ACM International Conference on Software Engineering and Information Management (ICSIM), Yokohama, Japan, 21–23 January 2022; pp. 113–117.
32. Norouzi, M.; Fleet, D.J.; Salakhutdinov, R.R. Hamming Distance Metric Learning. Advances in Neural Information Processing Systems 25 (NIPS 2012), 2012; Volume 25. Available online: https://proceedings.neurips.cc/paper_files/paper/2012/hash/59b90e1005a220e2ebc542eb9d950b1e-Abstract.html (accessed on 30 March 2024).
33. Gore, R.J.; Diallo, S.; Padilla, J. You are what you tweet: Connecting the geographic variation in America's obesity rate to twitter content. *PLoS ONE* **2015**, *10*, e0133505. [CrossRef] [PubMed]