J Ambient Intell Human Comput
DOI 10.1007/s12652-017-0536-y

ORIGINAL RESEARCH

A wikipedia-based semantic relatedness framework for effective dimensions classification in online reputation management

M. Atif Qureshi[1] · Arjumand Younus[1] · Colm O'Riordan[2] · Gabriella Pasi[3]

Received: 14 March 2017 / Accepted: 22 June 2017

Abstract Social media repositories serve as a significant source of evidence when extracting information related to the reputation of a particular entity (e.g., a particular politician, singer or company). Reputation management experts manually mine the social media repositories (in particular Twitter) for monitoring the reputation of a particular entity. Recently, the online reputation management evaluation campaign known as RepLab at CLEF has turned attention to devising computational methods for facilitating reputation management experts. A quite significant research challenge related to the above issue is to classify the reputation dimension of tweets with respect to entity names. More specifically, finding various aspects of a brand's reputation is an important task which can help companies in monitoring areas of their strengths and weaknesses in an effective manner. To address this issue in this paper we use dominant Wikipedia categories related to a reputation dimension; the dominant Wikipedia categories are then utilised within a semantic relatedness scoring framework to generate "associativities" with respect to the various reputation dimensions, and another version of "associativity" normalized by the "content entropy" of Wikipedia categories. The Wikipedia categories obtained through our applied methods are finally used in a random forest classifier for the task of reputation dimensions classification. The experimental evaluations show a significant improvement over the baseline accuracy.

Keywords Online reputation management · Semantic relatedness · Wikipedia · Reputation dimensions

1 Introduction

The area of "reputation management", emanating from the domain of "public relations", is concerned with managing the influence of an individual's or business's reputation (Fombrun and Shanley 1990). Studies have concluded that it is a driving force behind Fortune 500 corporate public relations since the beginning of the 21st century (Hutton et al. 2001). It essentially comprises (1) monitoring the reputation of an entity,[1] and (2) addressing content potentially damaging to the reputation of an entity.

With the growing popularity of social media the meaning of reputation management has shifted to online portals such as blogs, forums, opinion sites, and social networks. Companies are increasingly making use of social media for their promotion and marketing. At the same time social media users voice their opinions about various entities/brands (e.g., musicians, movies, companies) (Dellarocas

✉ M. Atif Qureshi
muhammad.qureshi@ucd.ie

Arjumand Younus
arjumand.younus@ucd.ie

Colm O'Riordan
colm.oriordan@nuigalway.ie

Gabriella Pasi
pasi@disco.unimib.it

[1] Insight Centre for Data Analytics, University College Dublin, Dublin, Ireland

[2] Information Technology Building, National University of Ireland, Galway, Ireland

[3] Dipartimento di Informatica, Sistemistica e Comunicazione, Universita Degli Studi Di Milano-Bicocca, Milano, Italy

[1] In the context of reputation management, an entity may refer to a celebrity, company, organization or brand.

Published online: 10 July 2017

Table 1 Description of reputation dimensions of an entity

Dimension	Description
Products and services	Products and services offered by the company or reflecting the consumers' satisfaction
Innovation	Innovativeness shown by the company, nurturing novel ideas and incorporating them into products
Workplace	Employees' satisfaction or the company's ability to attract, form and keep talented and highly qualified people
Citizenship	Company acknowledgement of community and environmental responsibility, including ethical aspects of the business: integrity, transparency, and accountability
Governance	The relationship between the company and the public authorities
Leadership	The leading position of the company
Performance	The company's long term business success and financial soundness

et al. 2003; Glance et al. 2005). This has recently given birth to "online reputation management" within the marketing domain where automated and semi-automated methods facilitate monitoring reputation of entities instead of relying completely on the manual reputation management by an expert (or a group of experts) as was traditionally done. Twitter serves as the most popular social media source for online reputation management (Jansen et al. 2009) due to its nature of enabling fast dissemination of information.

An entity has various *aspects* or *dimensions* that affect its reputation, and understanding these is a crucial step within "online reputation management". As an example, consider the following scenarios:

- A smartphone company releasing a new phone and creating hype around the product release.
- A pharmaceutical company in trouble due to release of a new drug without adequate testing.

In the first example above, the company's "products/services" are under discussion while in the second example the company's *governance* aspect is being examined.

As is obvious from above examples, automatic classification of tweets into various reputation dimensions is challenging on account of the lack of (1) context within tweets, and (2) explicit mention of terms that can impact an entity's reputation. In view of these challenges, we propose the incorporation of contextual knowledge from within an external source of evidence in order to solve the problem of reputation dimensions' classification. We utilize Wikipedia as the external source of evidence; our choice is motivated by its extensive coverage in the form of an effective hierarchy of categories and articles as explained in subsequent sections. Wikipedia categories are organized in a taxonomical manner serving as semantic tags for Wikipedia articles and this provides a strong abstraction and expressive mode of knowledge representation.

The remainder of this paper is organized as follows. In Sect. 2, we present relevant background for the undertaken task along with details of Wikipedia category-article structure. In Sect. 3, we discuss works related to this contribution while also explaining how our approach differs from existing ones. In Sect. 4, we explain the semantic relatedness framework which is essentially a core element of our methodology followed by a description of the main methodology within the employed machine learning approach that classifies tweets with respect to various reputation dimensions. In Sect. 5, we present the experimental evaluations. In Sect. 6, we conclude the paper with a discussion of possible future research directions.

2 Background

In this section we first present an overview of the reputation dimensions' classification task followed by a description of Wikipedia's key features particularly useful for the undertaken task.

2.1 Reputation dimensions' classification task

The task under involves classification of tweets according to the reputation dimensions which requires identification of various aspects significant to a company's reputation and Table 1 shows the standard dimensions used.[2] Basically, the task involves multi-class classification where given a tweet about an entity of interest and a set of reputation dimensions (in this case the ones shown in Table 1), the goal is to automatically classify the tweet to the single reputation dimension that the tweet relates.

2.2 Wikipedia

Wikipedia is a multilingual,[3] collaboratively constructed largest free encyclopedia containing over 4.4 million

[2] Note that these are the standard dimensions provided by the Reputation Institute.

[3] Available in 270+ languages.

articles[4] in English alone. Wikipedia contains articles on a wide range of topics from politics to science, news events to contributions by different people. Research have shown that Wikipedia is reasonably accurate[5] (Clauson et al. 2008) and as accurate as its rival commercial alternates i.e., Encyclopedia Britannica (Giles 2005) and Encarta (Rosenzweig 2006). We utilise Wikipedia on account of its rich category graph structure; and in order to enable exploitation of the Wikipedia information we develop our own system called WikiMadeEasy (Qureshi 2015).

A key difference between various knowledge bases lies in their underlying processing mechanism in terms of how they are read i.e., there exist human-readable and machine-readable knowledge bases. Wikipedia is different from other knowledge bases in terms of being human-readable.

Our main motivations behind use of Wikipedia are as follows:[6]

- Wikipedia is a collaboratively constructed resource which is updated extensively and hence, contains fresh knowledge on most topics.
- The continuous growth over a period of years makes it likely to stay useful over a number of years to come.
- The nature of continuous expansion of Wikipedia has made it truly the de-facto online encyclopedia which is more likely to cover aspects of human knowledge which are uncovered as of now but likely to be covered in future.
- Other knowledge bases chiefly rely on Wikipedia as potential source of knowledge while other sources are only included when Wikipedia lacks to cover them but this gap is more likely to diminish over the passing of time.

Each Wikipedia article contains content that defines a particular concept textually which may be accompanied with images related to the concept inside a Wikipedia page. Each article has a title that identifies a concept and each article can also be identified with zero or many redirect strings e.g., an article with title 'United States' can be identified by either its title or redirects such as 'USA' or 'US'. Furthermore, there is a possibility of ambiguity among different article titles, e.g., apple can either be a fruit or a company and likewise more than one person can have same names such as 'Michael Jordan' which can refer to the basketball

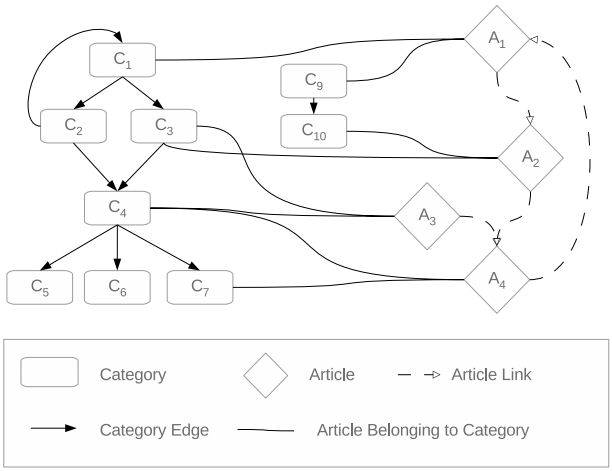

Fig. 1 Wikipedia category graph structure along with wikipedia articles

star in NBA or to the Professor at the University of California, Berkeley. To handle such ambiguous needs, Wikipedia has special pages which are called disambiguation pages. The disambiguation pages are special Wikipedia pages that contain one to many relations for ambiguous strings, e.g., the disambiguation page for 'apple' contains references to possible senses such as 'Apple (fruit)', 'Apple Inc. (company)', 'The Apple (1980 film)', etc. The Wikipedia articles are densely inter-connected to each other and each Wikipedia article references on average 22 other articles (Milne 2010). Furthermore, each article is mentioned inside different Wikipedia categories and each Wikipedia category generally contains parent and children categories.

Wikipedia categories are organized into a taxonomy structure (see Fig. 1). Each Wikipedia category can have an arbitrary number of subcategories as well as being mentioned inside an arbitrary number of supercategories (e.g., category C_4 in Fig. 1 is a subcategory of C_2 and C_3, and a supercategory of C_5, C_6 and C_7). Furthermore, in Wikipedia each article can belong to an arbitrary number of categories. As an example, in Fig. 1, article A_1 belongs to categories C_1 and C_9, article A_2 belongs to categories C_3 and C_{10}, while article A_3 belongs to categories C_3 and C_4. In addition to links between Wikipedia categories and Wikipedia articles, there are also links between Wikipedia articles as the dotted lines in Fig. 1 show (e.g., article A_1 outlinks to A_2 and has an inlink from A_4). The Wikipedia categories serve as a semantic tag for the articles to which they link (Zesch and Gurevych 2007). Figure 2 shows an existing Wikipedia category-article for the concept "Apple Inc."; note that the inlinks and outlinks between Wikipedia articles are organized according to the semantics inside the articles' content (e.g., the article on "Apple Inc." has

[4] http://en.wikipedia.org/wiki/Wikipedia:Size_of_Wikipedia.

[5] http://en.wikipedia.org/wiki/Reliability_of_Wikipedia.

[6] Note that we have essentially utilised the dumps made available by DBPedia. However, despite the fact that DBPedia contains a notable work of semantic annotations, we are not using this additional information.

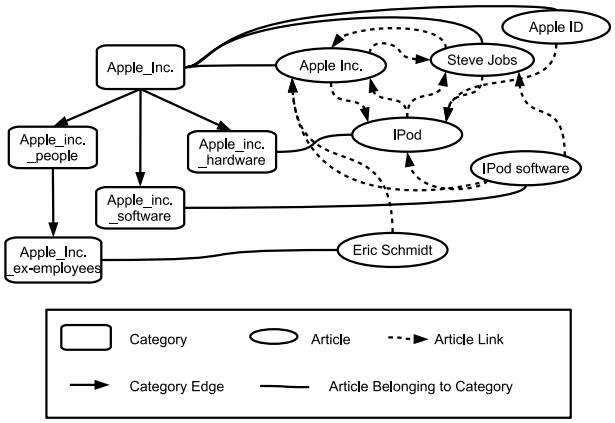

Fig. 2 Truncated category-article structure for concept "Apple Inc."

an inlink from the article on "Steve Jobs" while having an outlink the to article on "iPhone").

3 Related work

Our work touches the fields of semantic relatedness and text classification. We provide a brief overview of research that aims to associate textual data with their semantics and therefore, we begin by presenting the notion of "semantic relatedness" together with works on "semantic annotation". This is followed by covering some research works in the domain of "online reputation management".

3.1 Semantic relatedness

The literature has defined semantic relatedness as a means to allow computers to reason about written text (Witten and Milne 2008) whereby the reasoning deals with finding and quantifying the strength of semantic association between textual units (Hassan and Mihalcea 2011). Within the proposed works in the literature the difference lies in the knowledge base employed, the technique used for measurement of semantic distances and the application domain (Leal et al. 2012; Passant 2010).

We follow the notion of semantic relatedness adopted by Witten and Milne (2008) whereby we use it for measuring degree of similarity, and the relationship between different terms. Two examples from Witten and Milne (2008) are with respect to relationship between "social networks" and "privacy", and "cars" and "global warming". We however differ with them in terms of strategy employed since they utilise Wikipedia hyperlinks whereas our technique utilises Wikipedia categories in conjunction with Wikipedia articles. To estimate semantic relatedness, both Strube and Ponzetto (2006) and Gabrilovich and Markovitch (2007) used rich encyclopedic knowledge of Wikipedia. Strube and Ponzetto (2006) made a system called WikiRelate! which calculates the relatedness score of words by finding Wikipedia articles that contain words in their titles. They made use of previously developed measures for WordNet which in their calculation relied on the content of Wikipedia articles and the path distances found along the category taxonomy of Wikipedia. Gabrilovich and Markovitch (2007) proposed a technique called Explicit Semantic Analysis (ESA) which calculates Semantic Relatedness between words and text of any length [unlike (Strube and Ponzetto 2006) which operates over words only]; the technique bases itself on the vector space model using Wikipedia. The input is represented as a vector and is then scored on the basis of association with documents in the collection i.e., Wikipedia. Even though ESA gathered attention in the research literature (Gabrilovich and Markovitch 2009) it does not exploit the hypergraph of Wikipedia and this was filled by two later approaches (Witten and Milne 2008; Yeh et al. 2009). Witten and Milne (2008) made use of tf.idf-like measures on Wikipedia links and Yeh et al. (2009) made use of random walk algorithm [Personalized PageRank (Haveliwala 2002)] over the graph driven from Wikipedia's hyperlink structure, infoboxes, and categories.

3.2 Semantic annotation

The vision of the Semantic Web argues for a Web where Web resources are annotated with semantic metadata. In order to realize this vision, various works have focused on automatic approaches to perform semantic annotation. Earliest approaches within this area utilise machine learning over information extraction rules (Handschuh et al. 2002; Vargas-Vera et al. 2002). Other approaches make use of pattern-based methods which fundamentally rely on natural language processing techniques (Kiryakov et al. 2004; Laclavik et al. 2012). More recently, approaches utilising formal concept analysis have been proposed for semantic annotation of Web content (De Maio et al. 2014), and it provides a powerful representation through ordered lattics which is able to capture dependences among the concepts. Another line of work which is quite similar to semantic annotation is entity linking which comprises the process of enriching text with links to encyclopedic knowledge (chiefly, Wikipedia). The work by Mihalcea and Csomai (2007) pioneered sentence wikification by making use of Wikipedia in two independent processes of keyword extraction and word sense disambiguation; and finally, linking the disambiguated word sense to the correct Wikipedia article. Few semantic analysis/annotation works have been defined on top of sentence wikification such as document clustering (Hu et al. 2009), and content summarization (De Maio et al. 2016; Miao and Li 2010): it is seen that wikification enhances performance of the task at hand.

We fundamentally borrow the same ideas as that of wikification by making use of matching between Wikipedia categories and articles corresponding to various textual concepts. However, the major difference lies in the nature of the task where we explicitly consider reputation dimensions as explained in Sect. 2.

3.3 Online reputation management

Online reputation management which is a research area emanating from the marketing domain. Recent years saw initiatives and campaigns organized for addressing various computational challenges within this domain with the most notable being CLEF RepLab tasks (Amigó et al. 2013, 2014). Within these campaigns researchers proposed techniques to disambiguate entity names known, monitoring topics potentially damaging to a company's reputation, identifying dimensions/aspects significant to a company's reputation, and determining the type of author of Twitter profiles while also ranking Twitter authors by influence. This contribution focuses on the task of identifying various aspects central to the reputation of an entity, and of these seven aspects (listed in Table 1) are relevant in our context. Of the techniques proposed within CLEF RepLab evaluation campaign most relied on textual content from within tweets (Amigó et al. 2014) and more recently work by McDonald et al. (2015) investigates pseudo-relevant term expansion by means of a contemporary external Web corpus.

4 Methodology

In this section we first present the proposed semantic relatedness framework which constitutes the core of the methodology. This is followed by an explanation of the methodology employed to obtain dominant Wikipedia categories which are used within the semantic relatedness framework for extraction of machine-learning features useful for the reputation dimensions task.

4.1 Semantic relatedness based on wikipedia category-article structure

We follow the notion of semantic relatedness adopted by Witten and Milne (2008) whereby we use it as a means for inference of a relationship between textual units. Two examples from Milne and Witten are with respect to relationship between "social networks" and "privacy", and "cars" and "global warming". We model semantic relatedness as explicit and implicit connections between the concepts representing textual units and unlike previous works on semantic relatedness, our notion of semantic relatedness

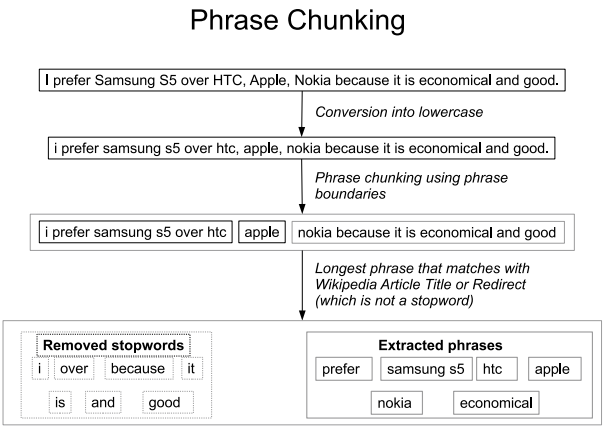

Fig. 3 Strategy of phrase chunking using Wikipedia

is not restricted to identification of relationships such as musician1:musician2[7] (Passant 2010) but can also identify relationships like microsoft:windows10.[8]

In the following sections, we first explain the process of candidate phrase generation performed through the chunking of textual data into variable-length phrases using Wikipedia. This is followed by an explanation of the strategy to produce relatedness scores through the exploitation of the Wikipedia category-article structure.

4.1.1 Generation of candidate phrases

Candidate phrases in the context of this contribution are the textual units extracted from the tweets, and we calculate semantic relatedness between these phrases and the pre-defined entity.[9]

Variable-length phrase chunking Figure 3 shows the phrase chunking strategy that we employ. In the first step, the textual content (i.e., a tweet) is converted into lowercase (to avoid case-sensitivity). Then, phrase boundaries (such as commas, semi-colons, sentence terminators etc.) are used for chunking the content into phrases. In the case of tweets, phrase boundaries also include tweet-specific markers (such as @, RT etc.). Finally, the extracted phrases are further reduced to those that match a Wikipedia article title or redirect. Preference is given to the extraction of the longest phrase. In the final step, there is an exception rule to ignore a phrase or word which matches exactly a

[7] musician1 and musician2 are two different musicians such as Madonna and Lady Gaga.

[8] Microsoft is a company whereas Windows10 is a product of Microsoft.

[9] It is this pre-defined entity corresponding to which reputation dimensions classification for the tweet has to be performed.

Table 2 Conventions

Convention	Explanation
RC	Set of parent category and subcategories to depth of 2 (i.e., list of categories in a hierarchy)
$Articles_{RC}$	Set of Wikipedia articles which are mentioned in at least one category from *RC*
WC	Set of all Wikipedia categories that mention Wikipedia articles in $Articles_{RC}$, therefore $RC \subset WC$

stopword. Figure 3 shows the removal of stopwords such as 'i', 'over', etc, and it also shows extracted phrases such as 'samsung s5', 'htc', etc.

We devise our strategy for variable-length phrase chunking by making two intuitive assumptions as follows:

- A phrase that contains more words is usually more informative than a phrase that contains less words, e.g., 'computer science' is more informative than 'science'.
- A single term which is not a stopword is more informative than a single term which is a stopword, e.g., 'science' is more informative than the stopword 'of'.

Note that we do not consider Wikipedia Miner (Milne and Witten 2013) for phrase chunking on account of its reliance on machine-learned approaches to disambiguating ambiguous terms. Our approach for phrase chunking requires a lightweight and unsupervised solution due to nature of Twitter volume.

4.1.2 Generation of relatedness scores

Note that the relatedness scores are generated for textual phrases (i.e. candidate phrases as explained in Sect. 4.1.1) with respect to a certain entity where an entity is a thing or concept with an independent existence such as a brand, company, celebrity, topical interest etc. For example, our aim can be to measure the relatedness of a piece of text to some real-world entity. Having extracted phrases from the text, we wish to score these phrases in terms of relatedness. In order to do so, we exploit the Wikipedia category taxonomies and the articles that are mentioned inside those category taxonomies as explained in the following subsection.

Wikipedia contains a huge and diverse amount of semantics pertaining to all entities in the form of related terms, article redirects, article hyperlinks, infoboxes,[10] parent and child categories etc. Wikipedia categories (i.e., parent and child categories) are particularly useful in that they can be used to infer or derive additional information pertaining to an entity. In fact, the Wikipedia category taxonomy can be representative of an entity; note that the choice of chosen category taxonomies to represent an entity is dependent upon the application scenario and we separately explain this process in Sect. 4.2. Here, for the sake of simplicity, we assume that a category taxonomy for which the relatedness score is to be calculated is an arbitrary category along with its sub-categories to a depth of two.[11] A depth of two is utilised as the optimal setting for inference of a relationship between a candidate phrase and Wikipedia category. Going further down in the depth is computationally expensive due to heavy interlinking of Wikipedia categories, whereas with a depth count of two we observed reasonable evaluation results without degrading performance. The inter-connections between the Wikipedia categories and Wikipedia articles are utilised in our semantic relatedness framework as explained below. Semantics is a broad term mainly used to represent the meaning and useful connections behind entities which is normally built upon extensive knowledge pertaining to an entity. As an example, the entity "Steve Jobs" represents the founder of company "Apple Inc."; however, to make this connection about entity "Steve Jobs" one would have to posses knowledge about entity "Apple Inc.".[12]

Each category taxonomy has exactly one parent category and usually several sub-categories. We refer to all these categories as *RC* (i.e., it contains all related categories in a hierarchy from depth count of zero to two). Note that the selection of *RC* by default includes all Wikipedia categories corresponding to a certain candidate phrase i.e., Wikipedia article, and according to application scenario certain Wikipedia categories are excluded from *RC*. In the case of reputation dimensions classification, our choice of *RC* is motivated by Wikipedia categories dominant in reputation dimensions, we explain the process for selection of *RC* in detail in Sect. 4.2. These categories *RC* contain different Wikipedia articles, we refer to these articles as $Articles_{RC}$. These articles $Articles_{RC}$ are also mentioned in categories other than *RC* and we retrieve all categories that contain

[10] An infobox is a fixed-format table designed to be added to the top right-hand corner of Wikipedia articles to consistently present a summary of some unifying aspect pertaining to the articles.

[11] It is important to note that a category representative of the entity is selected at this phase.

[12] An example category taxonomy for Apple Inc. can be seen on left side of Fig. 2.

$Articles_{RC}$ and refer to them as *WC*; note[13] that *RC* is a subset of *WC*. Table 2 summarizes the above-explained conventions. Note that in Fig. 2 all the Wikipedia categories that are shown (using the rounded rectangle symbol) represent *RC* and all the Wikipedia articles that are shown (using the oval symbol) represent $Articles_{RC}$.

The candidate phrases extracted from phrase chunking (explained in the previous section) that match an article title or redirect in $Articles_{RC}$ are called matched phrases. We use these matched phrases to calculate the relatedness score. In the next section, we summarize the factors which contribute in calculating the relatedness score of a candidate phrase using the Wikipedia category-article structure.

Relatedness measures The relatedness measures we propose aim to capture the closeness between two concepts within the Wikipedia category-article structure via the relatedness measure related to depth significance, and the number of common categories between two concepts via the relatedness measure related to category significance. Moreover, the significance of the phrase itself is taken into account so as not to overemphasize relatedness when the phrase itself is insignificant. We also present the aggregation of these measures into a single measure. In the formulations presented below we use the notation of (1) p to denote the candidate phrase for which a relatedness measure is to be calculated, and (2) cat_t to denote the category taxonomy corresponding to the entity under consideration.

We however propose a novel measure that takes into account depth at which a Wikipedia category matches an associated candidate phrase; note that it is the Wikipedia category representing our entity of interest and the candidate phrase of Sect. 4.1.1 for which relatedness is to be calculated. Below, we discuss three separate relatedness measures; these relate to depth, number of categories, and phrase frequency. Finally, we present the aggregation of these measures into a single measure. Note that we use non-normalized versions of relatedness measures as the range of values for Wikipedia category-article based heuristics is not wide, and moreover, we wish to capture even subtle relationships between the concepts represented by the textual units.[14]

$Depth_{significance}$ denotes the significance of category depth at which a matched phrase occurs. The underlying intuition behind this measure is that the deeper a match occurs in the taxonomy the less its significance to the entity under consideration. This means that the matched phrases in the parent category of the entity under investigation are more likely to be relevant to the entity than those at depth of two.

Each potential branch in a category is of a certain depth; the further down the category the greater is the specialization. As we move further down the category, we are potentially moving further away from the context expressed in the original subcategory (e.g., automata ⊂ computer science ⊂ science ⊂ knowledge).

$$Depth_{significance}(p, cat_t) = \sum_{cat \in RC \cap p_{categories}} \frac{1}{depth_{cat} + 1} \quad (1)$$

In the above formula, $p_{categories}$ denotes the categories in which the matched phrase appears. A $Depth_{significance}$ score is computed for each $p_{category}$ in *RC*, and an overall score for the considered matched phrase is obtained by summing up all the obtained significance scores. For an intuitive understanding of the $Depth_{significance}$ score, consider the Wikipedia article "Eric Schmidt" belonging to Wikipedia category "Apple Inc. ex-Employees" (refer to Fig. 2); the phrase "Eric Schmidt" is not highly related with the entity "Apple Inc." and this is also signified by its match with a Wikipedia category deeper in the hierarchy and hence, our formulation for $Depth_{significance}$ in above equation assigns a lower score to this phrase.

4.1.3 Heuristic 2: $Cat_{significance}$

$Cat_{significance}$ denotes the significance of the matched phrase as expressed by the number of categories containing it. Intuitively, a matched phrase is more related to an entity when the Wikipedia categories of a matched phrase coincide with the categories in the category taxonomy of the considered entity. Therefore, the more categories of a matched phrase in *RC*, the higher the significance of that particular matched phrase with respect to the entity.

$$Cat_{significance}(p, cat_t) = \frac{|RC \cap p_{categories}|}{|WC \cap p_{categories}|} \times log(|RC \cap p_{categories}| + 1) \quad (2)$$

$Cat_{significance}$ in the semantic relatedness model rewards the matched phrases which are densely inter-connected within the categories in *RC*.

4.1.4 Heuristic 3: $Phrase_{significance}$

$Phrase_{significance}$ is a combination of phrase word length and frequency of the phrase within the textual block from where

[13] E.g., Wikipedia article "Steve Jobs" of "Apple Inc." is mentioned inside a category "1955 births" which is not present either in parent nor in sub-categories of entity's Wikipedia article.

[14] Normalizing a subtle relationship may result into mathematical zero due to small fraction and storing a low fraction with high precision is not an efficient choice.

it's extracted.[15] Intuitively, the greater the phrase length,[16] the more informative or important it becomes, likewise the more frequent the phrase is in the textual block from where it's extracted, the more importance it assumes.

$$Phrase_{significance}(p, cat_t) = log(wordlen(p) + 1) \times p_{frequency} \tag{3}$$

$$p_{frequency} = log(freq + 1) \tag{4}$$

We combine the three separate relatedness scores of $Depth_{significance}$, $Cat_{significance}$, and $Phrase_{significance}$ to give a unique relatedness score. More than one approach is possible for the aggregation of these measures, however we adopt[17] the following.

$$Relatedness(p, cat_t) = Depth_{significance}(p, cat_t) \times Cat_{significance}(p, cat_t) \times Phrase_{significance}(p) \tag{5}$$

So far we have discussed generation of relatedness scores for matched phrases. In this paper these matched phrases are essentially taken from a tweet. The combined effect of $Depth_{significance}$, $Cat_{significance}$, $Phrase_{significance}$, and combined *Relatedness* is applied over the entire tweet via the following summations:

$$Depth_{significance}(tweet, cat_t) = \sum_{p \in MatchedPhrases} Depth_{significance}(p, cat_t) \tag{6}$$

$$Cat_{significance}(tweet, cat_t) = \sum_{p \in MatchedPhrases} Cat_{significance}(p, cat_t) \tag{7}$$

$$Phrase_{significance}(tweet) = \sum_{p \in MatchedPhrases} Phrase_{significance}(p) \tag{8}$$

$$Relatedness(tweet, cat_t) = \sum_{p \in MatchedPhrases} Relatedness(p, cat_t) \tag{9}$$

Here, *MatchedPhrases* is used to denote the set of matched phrases that occur in a given tweet in Eqs. 6–9.

4.2 Multi-class classification into reputation dimensions

Recall from Sect. 2.1 that the reputation dimensions classification task requires multi-class classification of tweets into pre-defined classes that reflect which aspect of an entity's reputation is under discussion. Again, Table 1 shows the standard dimensions used. In the subsections that follow we present an explanation of the methodology to extract dominant Wikipedia categories for utilisation within the semantic relatedness framework. This is followed by a brief explanation of the feature set used for the reputation dimensions classification task.

4.2.1 Extraction of dominant wikipedia categories

As we noted in Sect. 4.1 our framework requires preselected Wikipedia categories representative of the entity under investigation and this choice is motivated by the application scenario. Herein, we describe the process through which we select Wikipedia categories for the reputation dimensions classification task.

Using the training data we select the top category taxonomies by first combining the training tweets of a single reputation dimension into one document, and then we perform the process of variable-length phrase chunking (as explained in Sect. 4.1.1) to extract candidate phrases. Each matched Wikipedia article corresponding to a candidate phrase[18] belongs to one or more Wikipedia categories, and we label this set of categories as *WikiCategories*. From this training data, we maintain a voting count corresponding to each Wikipedia category (i.e., $WC_{training}$) through which the strength of association of a Wikipedia category (i.e., $Associativity_{WC_{training}}$) with respect to a reputation dimension is calculated as follows.

Using these Wikipedia categories we determine the strength of association of a Wikipedia category (i.e., $Associativity_{WC_{training}}$) with respect to a reputation dimension as follows:

$$Associativity_{WC_{training}} = \frac{n_i(WC_{training})}{\sum_{i \in WikiCategories} n_i(WC_{training})} \tag{10}$$

Equation 10 essentially represents how often a certain Wikipedia category, $WC_{training}$, occurs in labelled tweets[19] normalized by all other occurrences of Wikipedia categories[20] corresponding to a particular reputation dimension. In this way it models the strength of association of a given Wikipedia category (i.e., $WC_{training}$) with the reputation dimension under consideration.

$Associativity_{WC_{training}}$ fails to take into account the effect of noise in tweets corresponding to the training data of a

[15] This could be a paragraph, sentence or tweet.

[16] Number of words in a phrase.

[17] Empirically this aggregation performs reasonably well during the evaluations as shown in the later chapters.

[18] Recall from Sect. 4.1.1 that the final step in extraction of candidate phrases corresponds to matching with Wikipedia article titles.

[19] From within training data.

[20] From the set *WikiCategories* that represents all Wikipedia categories within a given reputation dimension.

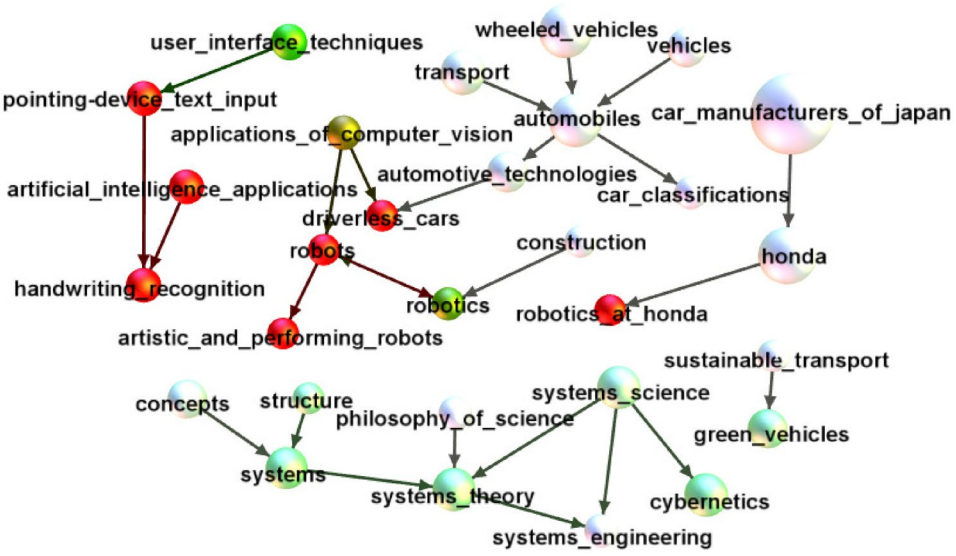

Fig. 4 Wikipedia categories for reputation dimension "Innovation" (from training data) for automotive domain

certain reputation dimension, and this noise may potentially affect the choice of Wikipedia category taxonomies that are selected for the semantic relatedness framework. To alleviate this, we introduce a measure that represents the uncertainty associated with a certain Wikipedia category using $ContentEntropy_{WC_{training}}$ defined as follows:

$$ContentEntropy_{WC_{training}} = -\sum_{i \in RD} p_i(WC_{training}) log(p_i(WC_{training})) \tag{11}$$

$$p_i(WC_{training}) = \frac{n_i(WC_{training})}{\sum_{i \in RD} n_i(WC_{training})} \tag{12}$$

As Eq. (12) shows $p_i(WC_{training})$ represents the proportion of times a certain Wikipedia category occurs in tweets of a particular reputation dimension as opposed to all reputation dimensions.[21]

We finally introduce a modified version of the associativity measure known as $RelativeAssociativity_{WC_{training}}$:

$$RelativeAssociativity_{WC_{training}} = \frac{Associativity_{WC_{training}}}{ContentEntropy_{WC_{training}}} \tag{13}$$

To aid the reader in visualizing the dominant Wikipedia categories, we plot the obtained categories using Gephi[22] whereby associativity scores are plotted to select the Wikipedia categories most closely related to a given reputation dimension. Figure 4 illustrates the graph of Wikipedia categories corresponding to the reputation dimension of "Innovation" for the automotive domain. The red-colored nodes in this figure represent the Wikipedia categories that occur in a particular dimension with high associativity scores, the white-colored nodes represent low associativity scores, and the various green-colored nodes represent moderate associativity scores.

4.2.2 *Set of features based on wikipedia category-article structure*

Using the category taxonomies representing the highest associativity scores as described in Sect. 4.2.1, we generate the feature set based on Wikipedia category-article structure. For each category taxonomy we generate a score corresponding to *Depth Significance* (i.e., Eq. 6), *Category Significance* (i.e., Eq. 7), and finally *Relatedness* (i.e., Eq. 9) as the set of features.

5 Experimental evaluations

This section describes the experimental procedure that we undertake to demonstrate the effectiveness of the proposed methods. First, we present details of experimental data and environment and finally, we present the experimental results.

5.1 Dataset and environment

5.1.1 *Twitter dataset*

We use the dataset provided by CLEF 2014 RepLab task organizers which is a multi-lingual collection of tweets

[21] Note that *RD* represents the set of all seven reputation dimensions.

[22] http://gephi.github.io.

Table 3 Results of reputation dimensions' classification task of RepLab 2014

Approach	Accuracy	F-measure
McDonald_RD_1	0.6073	0.3195
DAE_RD_1	0.7231	0.3906
Lys_RD_1	0.7167	0.4774
SIBTEX_RD_1	0.7073	0.4057
$Associativity_{WC}$	0.7615	0.5132
$RelativeAssociativity_{WC}$	0.7802	0.5509
Baseline	0.6222	0.4072

Table 4 Results of reputation dimensions' classification task of RepLab 2014

Approach	Precision	Recall	F-measure	Accuracy
RepLab Best	0.4928	0.4697	0.4810	0.7319
$Associativity_{WC}$	0.7819	0.3819	0.5132	0.7619
$RelativeAssociativity_{WC}$	0.8013	0.4197	0.5509	0.7802
McDonald	0.7502	0.3861	0.5016	0.7431
Gabrilovich and Markovitch	0.7604	0.3694	0.4973	0.7494

(i.e., 20.3% Spanish tweets and 79.7% English tweets). The corpus contains tweets referring to a set of 31 entities from two domains; automotive and banking. The tweets were gathered by organizers of the task by issuing the entity's name as the query. For each entity roughly 2300 tweets were collected with the first 750 constituting the training set, and the rest serving as the test set.

5.1.2 Wikipedia

The data for Wikipedia category-article structure is obtained through a custom Wikipedia API that has pre-indexed Wikipedia data and hence, it is computationally fast.[23] The API has been developed using the DBPedia (Bizer et al. 2009) dumps and it is a programmer-friendly API enabling developers and researchers to mine the huge amount of knowledge encoded within the Wikipedia structure.

5.2 Experimental setup

Using the feature sets described in Sect. 4.2, we train a random forest classifier over the training data and then use it to predict labels for the test data. We perform two machine learning runs as follows:

1. For the first run, we use Wikipedia categories generated by $Associativity_{WC}$ (i.e., Eq. 10) within the semantic relatedness framework for generation of features.
2. For the second run, we use Wikipedia categories generated by $RelativeAssociativity_{WC}$ (i.e., Eq. 13) within the semantic relatedness framework for generation of features.

In both settings, we train a random forest classifier per-domain i.e. combining all tweets related to a particular domain into one training and one test set.

5.3 Experimental results

Table 3 presents experimental results for the reputation dimensions classification task, where $Associativity_{WC}$ and $RelativeAssociativity_{WC}$ represent experimental runs explained above. As can be seen from Table 3, our approach to tackle the reputation dimensions classification task outperforms all other known methods within CLEF RepLab 2014 evaluation campaign. This demonstrates the effectiveness of Wikipedia category and article associations, and when used in conjunction with entropy scores for each reputation dimension the experimental results improve further.

Table 4 presents experimental outcomes when compared with previously best-known techniques in the literature. In particular our approach outperforms the tweet enrichment mechanism proposed by McDonald et al. (2015). Moreover, in order to provide a deeper insight into the power of Wikipedia category-article we also perform an experimental run by enriching tweet vectors through Explicit Semantic Analysis (Gabrilovich and Markovitch 2007), and again our approach demonstrates superior performance.

6 Conclusion and future work

In this paper, we explored the effectiveness of Wikipedia's category-article structure via its application in the domains of tweet classification for online reputation management. More specifically, the relationship between Wikipedia categories and articles is explored via a textual phrase matching framework whereby the starting point is textual phrases (from within tweets) that match Wikipedia articles' titles/redirects. The Wikipedia articles for which a match occurs are then utilised by extraction of their associated categories, and these Wikipedia categories are used to derive various structural measures such as those relating to taxonomical depth and Wikipedia articles they contain. Furthermore,

[23] http://bit.ly/1eMADG9, we aim to release the API as an open source Wikipedia tool to facilitate other researchers.

the concept of "associativies" and "relative associativities" of Wikipedia categories refines the feature set used in tweet classification, and helps alleviate the problem of lack of context in tweets to a major extent. An interesting research direction worth exploring is utilisation of semantic relatedness in combination with traditional text similarity measures such as cosine similarity, jaccard similarity etc. to make stronger inferences from within textual data. This can help alleviate the limitation arising due to noise in Wikipedia category-article structure thereby assisting in addressing some limitations of the current methodology.

References

Amigó E, De Albornoz JC, Chugur I, Corujo A, Gonzalo J, Martín T, Meij E, De Rijke M, Spina D (2013) Overview of RepLab 2013: evaluating online reputation monitoring systems. In: Information access evaluation. Multilinguality, multimodality, and visualization, Springer, pp 333–352

Amigó E, Carrillo-de Albornoz J, Chugur I, Corujo A, Gonzalo J, Meij E, de Rijke M, Spina D (2014) Overview of RepLab 2014: author profiling and reputation dimensions for online reputation management. In: Information access evaluation. Multilinguality, multimodality, and interaction, Springer, pp 307–322

Bizer C, Lehmann J, Kobilarov G, Auer S, Becker C, Cyganiak R, Hellmann S (2009) Dbpedia—a crystallization point for the web of data. Web Semant Sci Serv Agents World Wide Web 7(3):154–165

Clauson KA, Polen HH, Boulos MNK, Dzenowagis JH (2008) Scope, completeness, and accuracy of drug information in wikipedia. Ann Pharmacother 42(12):1814–1821

De Maio C, Fenza G, Gallo M, Loia V, Senatore S (2014) Formal and relational concept analysis for fuzzy-based automatic semantic annotation. Appl Intell 40(1):154–177

De Maio C, Fenza G, Loia V, Parente M (2016) Time aware knowledge extraction for microblog summarization on twitter. Inf Fusion 28:60–74

Dellarocas C, Awad NF, Zhang XM (2003) Exploring the value of online reviews to organizations: implications for revenue forecasting and planning. Manag Sci 30:1407–1424

Fombrun C, Shanley M (1990) What's in a name? reputation building and corporate strategy. Acad Manag J 33(2):233–258

Gabrilovich E, Markovitch S (2007) Computing semantic relatedness using wikipedia-based explicit semantic analysis. IJCAI 7:1606–1611

Gabrilovich E, Markovitch S (2009) Wikipedia-based semantic interpretation for natural language processing. J Artif Intell Res 34(2):443

Giles J (2005) Internet encyclopaedias go head to head. Nature 438(7070):900–901

Glance N, Hurst M, Nigam K, Siegler M, Stockton R, Tomokiyo T (2005) Deriving marketing intelligence from online discussion. In: Proceedings of the eleventh ACM SIGKDD international conference on knowledge discovery in data mining, KDD '05, pp 419–428

Handschuh S, Staab S, Ciravegna F (2002) S-cream—semi-automatic creation of metadata. In: International conference on knowledge engineering and knowledge management, Springer, pp 358–372

Hassan S, Mihalcea R (2011) Semantic relatedness using salient semantic analysis. In: AAAI, pp 884–889

Haveliwala TH (2002) Topic-sensitive pagerank. In: Proceedings of the 11th international conference on World Wide Web, ACM, pp 517–526

Hu X, Zhang X, Lu C, Park EK, Zhou X (2009) Exploiting wikipedia as external knowledge for document clustering. In: Proceedings of the 15th ACM SIGKDD international conference on knowledge discovery and data mining, ACM, pp 389–396

Hutton JG, Goodman MB, Alexander JB, Genest CM (2001) Reputation management: the new face of corporate public relations? Public Relat Rev 27(3):247–261

Jansen BJ, Zhang M, Sobel K, Chowdury A (2009) Twitter power: tweets as electronic word of mouth. J Am Soc Inf Sci Technol 60(11):2169–2188

Kiryakov A, Popov B, Terziev I, Manov D, Ognyanoff D (2004) Semantic annotation, indexing, and retrieval. Web Semant Sci Serv Agents World Wide Web 2(1):49–79

Laclavik M, Šeleng M, Ciglan M, Hluchỳ L (2012) Ontea: platform for pattern based automated semantic annotation. Comput Inform 28(4):555–579

Leal JP, Rodrigues V, Queirós R (2012) Computing semantic relatedness using dbpedia. In: Symposium on languages, applications and technologies, 1st, Schloss Dagstuhl, pp 133–147

McDonald G, Deveaud R, McCreadie R, Macdonald C, Ounis I (2015) Tweet enrichment for effective dimensions classification in online reputation management. In: Ninth international AAAI conference on web and social media, Oxford

Miao Y, Li C (2010) Enhancing query-oriented summarization based on sentence wikification. In: Workshop of the 33rd annual international ACM SIGIR conference on research and development in information retrieval, Oxford, p 32

Mihalcea R, Csomai A (2007) Wikify!: linking documents to encyclopedic knowledge. In: Proceedings of the sixteenth ACM conference on conference on information and knowledge management, ACM, pp 233–242

Milne D, Witten IH (2013) An open-source toolkit for mining wikipedia. Artif Intell 194:222–239

Milne DN (2010) Applying wikipedia to interactive information retrieval. PhD thesis, University of Waikato

Passant A (2010) Measuring semantic distance on linking data and using it for resources recommendations. In: AAAI spring symposium: linked data meets artificial intelligence, vol 77, p 123

Qureshi MA (2015) Utilising wikipedia for text mining applications. PhD thesis, NUI, Galway, Ireland

Rosenzweig R (2006) Can history be open source? Wikipedia and the future of the past. J Am Hist 93(1):117–146

Strube M, Ponzetto SP (2006) Wikirelate! computing semantic relatedness using wikipedia. AAAI 6:1419–1424

Vargas-Vera M, Motta E, Domingue J, Lanzoni M, Stutt A, Ciravegna F (2002) Mnm: ontology driven semi-automatic and automatic support for semantic markup. In: International conference on knowledge engineering and knowledge management, Springer, pp 379–391

Witten I, Milne D (2008) An effective, low-cost measure of semantic relatedness obtained from wikipedia links. In: Proceeding of AAAI workshop on wikipedia and artificial intelligence: an evolving synergy. AAAI Press, Chicago, pp 25–30

Yeh E, Ramage D, Manning CD, Agirre E, Soroa A (2009) Wikiwalk: random walks on wikipedia for semantic relatedness. In: Proceedings of the 2009 workshop on graph-based methods for natural language processing, association for computational linguistics, pp 41–49

Zesch T, Gurevych I (2007) Analysis of the wikipedia category graph for NLP applications. In: Proceedings of the TextGraphs-2 workshop (NAACL-HLT), Oxford