Improving Document Clustering Performance: The Use of an Automatically Generated Ontology to Augment Document Representations

Stephen Bradshaw[1], Colm O'Riordan[1] and Daragh Bradshaw[2]
[1]*Information Technology, National University Ireland Galway, Galway, Ireland*
[2]*Department of Psychology, National University Limerick, Ireland*

Keywords: Document Clustering, Graph Theory, WordNet, Classification, Word Sense Disambiguation, Data Mining.

Abstract: Clustering documents is a common task in a range of information retrieval systems and applications. Many approaches for improving the clustering process have been proposed. One approach is the use of an ontology to better inform the classifier of word context, by expanding the items to be clustered. Wordnet is commonly cited as an appropriate source from which to draw the additional terms; however, it may not be sufficient to achieve strong performance. We have two aims in this paper: first, we show that the use of Wordnet may lead to suboptimal performance. This problem may be accentuated when a document set has been drawn from comments made in social forums; due to the unstructured nature of online conversations compared to standard document sets. Second, we propose a novel method which involves constructing a bespoke ontology that facilitates better clustering. We present a study of clustering applied to a sample of threads from a social forum and investigate the effectiveness of the application of these methods.

1 INTRODUCTION

Social media has become ubiquitous in our society and there are many examples of it being the catalyst for huge social change (Shirky, 2011). Examples of the use of social media as a social influence can be found in the use of twitter during the Arab Spring (Lotan et al., 2011); the Green Movement in Iran (Golkar, 2011); and generally in political elections, most notably in American presidential elections (Bennett, 2012).

We use data collected from Reddit to conduct our experiments. Reddit is the self titled *front page of the internet* because it represents posts that people have seen online and deem worthy of others' attention (Singer et al., 2014). Reddit has seen huge growth as a social platform since its creation in 2005, and the number and diversity of the posts has equally seen exponential growth (Singer et al., 2014). We expand more on the structure of Reddit in Section 2.

Clustering is the act of grouping documents of a similar nature together. It can be done in either overlapping or disjointed means. Overlapping allows for documents to occur in different clusters while disjointed allows for the allocation of one instance of a document per cluster. Clustering is an example of unsupervised learning as the clustering agent has no a priori knowledge of category labels. K-means is an approach to *partitioning* where a document is allocated to one of a fixed number of predetermined clusters. It is an iterative process which reassigns the documents to the cluster which are most similar to itself. The process continues until a predetermined termination point or when there is no more reassigning of documents to be done (Shah and Mahajan, 2012).

K-means is an established method that has been used since its creation in 1955. The longevity of K-means as well as the plethora of clustering approaches that have stemmed from it are a testament to its robustness. Like many clustering approaches however, there are a number of issues with using it: how one represents the documents being clustered, how to combat noise as a result of the high dimensional nature of documents and the need to know a priori how many clusters are required (Shah and Mahajan, 2012). More concretely, the presence of synonymy and polysemy in words means that a direct mapping of document to document may not accurately return documents of similar nature which use different terms to capture a concept. This phenomenon is noted by (Deerwester et al., 1990) who use statistical methods to improve recall in document queries. One proposed solution to reduce the noise associated with text documents is the use of an ontology to clarify the intended use of each word. Building on the principle of quantifying the intended meaning of a word through the use

Bradshaw S., O'Riordan C. and Bradshaw D.
Improving Document Clustering Performance: The Use of an Automatically Generated Ontology to Augment Document Representations.
DOI: 10.5220/0006500202150223
In *Proceedings of the 9th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (KDIR 2017)*, pages 215-223
ISBN: 978-989-758-271-4

Table 1: List of all threads and sizes after processing.

Thread_A	Thread_B	Thread_A_Size	Thread_B_Size	Total Comments
Rugbyunion	Quadcopters	926	672	1598
LearnPython	Worldnews	904	513	1417
Movies	Politics	422	963	1385
Music	Boardgames	448	942	1390
England	Ireland	743	935	1678

of an external ontology, we propose the construction of a bespoke lexicon.

A notable feature of Reddit is that there are many sub-domains similar in nature. This paper proposes using identified related domains as a source of evidence, and using graph principles to construct a bespoke ontology that can be used to better augment comments leading to more precise clustering of the original collection. Our graph approach allows for the creation of clusters that are not reliant on directly mapping terms with terms in documents, but instead will allow for concepts to be represented by a number of closely related terms. Our results show that such an approach considerably improves performance. Using a bespoke ontology is beneficial over a relatively general and static ontology such as Wordnet as better expansion candidate are provided. Upon clustering user comments, using a graph representation of the ontology reduces the ambiguity in the original comments through polysemy resolution. These two factors; a better ontology and a graph model, benefit the clustering through better document representation.

The paper outline is as follows: in the next section, we discuss Reddit, the source of data for our clustering experiments. In section 3, we outline our methodology before discussing related work in the subsequent section.

2 RELATED WORK

The original authors of Wordnet (Hotho et al., 2003) show that document clusters can be improved with the addition of background material. They use synonyms and hypernyms to augment their document vectors (Hotho et al., 2003). They investigated three approaches; firstly **All Concepts** which involves taking all of the related terms and using these to augment the document. Secondly, **First Concept** which entails replacing the term in the document with the identified related term. Finally they used a **Disambiguation by Context Approach**, which involved using the definition of the term in question and measuring the similarity with the words found in the document.

There have been many subsequent papers that have shown how Wordnet can be used to improve clusters. Baghel et al (Baghel and Dhir, 2010) propose an algorithm (**Frequent Concept Based Document Clustering (FCBDC)**) which identifies frequently occurring concepts. They define concepts as words that have the same meaning and use WordNet to identify when words have a similar meaning. They subsequently appoint each concept as a kernel and cluster the documents around them. Their approach involves using *first concept* so the initial words are replaced with their synonyms.

Weng et al. (Wang and Hodges, 2006) investigate if the use of semantic relatedness can be used to cluster using Word Sense Disambiguation (WSD) principles. They define semantic relatedness as *a criterion to scale the relatedness of two senses in a semantic network* (Wang and Hodges, 2006). They used a Part of Speech Tagger (POS) to identify the grammatical use of each word. Each document was converted this way so that the words in the vector space model were converted into tuples of the term and the POS of the term. The POS were then used as the intended context of the word. They applied this approach to a corpus of 1600 abstracts of 200 words each. These abstracts were further divided into 8 different categories. They found that through this approach they were able to improve upon results however, the small scale of their experiment meant that results were hindered through lack of sufficient distinguishing features for each category.

Mahjan et al. (Mahajan and Shah, 2016) use an approach similar in nature to the work of Wang et al. (Wang and Hodges, 2006). They investigate ways in which one can improve upon cluster results. They investigate changing the number of clusters, the use and stop word removal and lemmatisation. In addition, they engaged with Wordnet to augment the document vectors with their respective synonyms. Using a POS tagger to identify if a word is a noun, verb, adjective they augmented the the vectors with the most closely identified one. Their approach was applied to a Reuters document corpus and they achieved an improvement of 11% for purity and 29 % for entropy in the 20 news group and additionally they got an improvement of 18% and 38% respectively on the Reuters corpus.

Another approach which utilises Wordnet to improve upon traditional clustering is the work of Hung et al. (Hung et al., 2004). They endeavour to better classify news articles as found in the Reuters text corpus. They take 200,000 articles with over 50 classifications. They use the Wordnet to identify if a hypernymy exists for a given word. They use a replace policy to reduce dimensionality (**First Concept**), whereby the hyponym will be replaced with its hypernym. Their approach allows for multiple classifications. Similar to how Latent Dirchlet Allocation (LDA) works they count the words in the document as well as per topic (Blei et al., 2002). As documents allow for multiple topics, words can be used to indicate the presence of different topics. As well as hosting hypernyms and synonyms, Wordnet contains a definition of each potential meaning of the target word as well as an example of usage. This is similar to the approach *Disambiguation by Context Approach* used by the original authors to disambiguate the intended meaning for the word used (Hotho et al., 2003).

The premise of the approach of Zheng et al. (Zheng et al., 2009) is that documents are made up of concepts and that different terms are often used to describe a concept. By resolving the various terms to their concepts one can improve upon information retrieval. The authors focus on noun phrases and the semantic patterns inherent in documents. Stemming and stopword removal are important preprocessing steps in achieving this end. They define a **noun phrase** as *a grammatical category (or phrase) which normally contains a noun as its head and which can be modified in many ways* (Zheng et al., 2009). The authors propose using syntactical analysis to identify the noun phrases. Partial parsing is used, which means that the appropriate noun phrases are analysed rather than the document as a whole. They analyse noun phrases by considering the synonyms of the adjective. Then the relationship between the noun phrases are explored. Synsets are a useful tool here as the first synset is the most common for of the term. Additionally WordNet is used to identify the hypernyms, hyponyms, meronyms and holonyms which are resolved to one representative concept. The authors find the use of hypernyms to be most effective; they speculated that the use of hypernyms is most effective because *document categorization tends to more naturally on the more-general terms rather than more-specific terms* (Zheng et al., 2009).

The approach of this paper of mapping semantic similarity by considering re-occurring proximity of terms is similar in nature to the approach proposed by Lund and Burgess (Lund and Burgess, 1996). They construct a dataset from Usenet and map the strength of the proximity of terms. They use a sliding window of 10 and store the terms in a matrix. The proceeding terms are stored row-wise and the preceding terms as stored in a column-wise fashion. They use multi-dimensional scaling to draw inferences on the associations of the target terms. They conduct three experiments and make the following conclusions. First, the euclidean distance of the associative matrices of terms show that proximity is related to the frequent co-occurrence of terms. Second, Categorical relationships can be ascertained through this approach. More concretely they show that hyponym terms can be grouped with their corresponding hypernyms. Third, that automatically determined semantic distance between terms are comparable with human judgment of the same.

Other work that embodies the Hyperspace Analogue to Language (HAL) approach is that of Song and Bruza (Bruza and Song, 2001). They aim to model the information flow that is created when two terms are located adjacent to one another. The concept those two terms represent is the sum of their associate terms. Applying HAL, they create a matrix of associative terms. Each concept they propose is the sum of those associative terms. They normailise those matrices values by taking the strength of the reoccurring terms and dividing it by all the terms, after the terms that have fallen below a quality threshold have been removed. $w_{c_ip_j} = \frac{w_{c_ip_j}}{\sqrt{\sum_k w_{c_ip_k^2}}}$. To find the strength of the common terms between two concepts they apply the formula $w_{c_1p_i} = l_1 + \frac{l_1 * w_{c_1p_i}}{\max_k(w_{c_1p_k})}$, where l_1 represents a weight that reflects that this particular concept is more dominant than the second concept. They make no statistical evaluation of their approach other than to discuss the relatedness of the vectors produced by this process.

3 DATASET

To conduct our investigation, data from Reddit was utilised. Social platforms like Reddit are forums where users create and curate the comments found at the site. The quality of the content is appraised through a voting system. Reddit has been described as a *Web-democracy* because everyone has a voice and can express their opinion (Weninger et al., 2013). Social media platforms mark a divergence from traditional media outlets where there are a handful of curators who dictate the conversation. Instead, everyone is free to suggest a topic and the impact of that suggestion is felt in the number of people who engage with the narrative. The hierarchical structure of the conversation threads allows for divergences in topics.

The structure of Reddit is as follows:

- **Subreddits:** These are sub domains of a common theme. They comprise users who have an interest in that theme. Related topics to that general theme are submitted by users and people then engage with those topics through making comments on the original posts. All subreddits have at least one moderator who ensures that the rules of that thread are upheld and that relevance is maintained.
- **Posts:** are the initial comment submissions made in a subreddit. It can be in the form of text, image or links. It is an invitation to other users to engage in a discourse or express their opinion on an issue through up-voting or down-voting.
- **Comments:** are user submissions to an initial post. They can be new comments (referred below as parent comments) which mark a new perspective on the post or they can be comments on existing comment threads (child comments), this indicates that the point is related to that branch of conversation.
- **Parent Comments:** refers to the original comment made to a post. Child comments are all of the subsequent comments posted to that comment thread. One post can have many parent comments and each parent comment can have a number of children comments.
- **Voting:** allows for each user to express their opinion on a post or comment. It is in the form of an upvote or a downvote. There is an algorithm that ranks the votes which informs where that post is placed in the thread hierarchy. A particular popular post can make it to the front page of Reddit which is not topic specific and will garner increased attention through increased visibility.
- **Karma:** represents the overall feedback that a user has on the user's collective posts. Each upvote is an additional karma point and each downvote takes from the users overall accrued karma points.

A user has the option to *subscribe* to a thread. Typically a user will be subscribed to a number of different threads, and the most popular posts in these are shown on the user's personal wall. In addition there is a universal wall which displays the most commented upon or upvoted posts and can come from any thread.

Table 2: List of all threads and and their related thread.

Thread	Related Thread
Rugbyunion	NRL
LearnPython	Python
Movies	Fullmoviesonyoutube
Music	popheads
England	London
Quadcopters	Quadcopter
Worldnews	News
Politics	ukpolitics
Boardgames	Risk
Ireland	Dublin

4 METHODOLOGY

4.1 Introduction

To perform our experiments in improving clustering of documents from social media forums and improving performance of said clustering, we first created a document collection. We selected 10 threads for the purpose of running our experiments and a further 10 threads for the creation of relevant ontologies; the threads used are listed in Table 2. As a baseline experiment, we apply K-means clustering to the content of the 10 threads. We then re-apply that clustering approach on the same documents using augmented versions of the documents. We analyse a number of document expansion techniques from the literature and finally our own approach. We now discuss how the dataset was constructed and processed, the baseline approach, how WordNet was used to augment the document set, and finally, we describe in detail our own proposed approach.

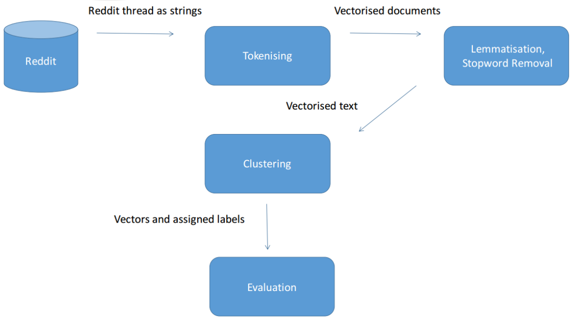

Figure 1: Steps taken for standard approach.

4.2 Baseline Approach

First we tokenise the text; this involves removing stopwords, lemmatising and vectorising. We use a K-means as implemented in Sklearn (Pedregosa et al.,

2011) to cluster the documents. For the basic approach we clustered the document set in it's vectorised form, and recorded the results. See Figure 1 for a graphical representation of the process.

Table 1 contains the names and sizes of each of our threads. To test the robustness of our approach we took each thread in Thread_A and combined it with each of the threads found in the Thread_B column. This produced a result set of 25 collections; we then attempt to cluster the documents back into the two original clusters. The results of this are discussed in the results section. We selected threads that were similar in nature, because we were attempting to capture the nuanced speech associated with a thread. We felt that if we could single out threads that would have common members, we would better be able to model the language used. All approaches discussed below were evaluated in the same manner. Each document representation was augmented in a different way according to the approach being investigated.

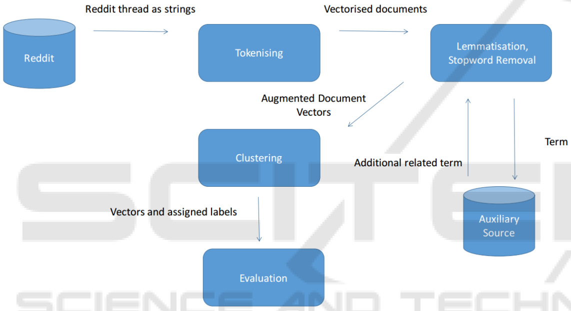

Figure 2: Steps taken for augmented approach.

4.3 Wordnet Approach

To measure the impact of including synonyms and hypernyms, Wordnet was used. Figure 2 shows the process. The initial steps are the same as the baseline. Each document is represented as a vector. Stopword removal and lemmatisation was then applied. Each document is augmented with related hypernyms. We took the corpus and checked if a word had a hypernym, and if so, that hypernym was added to the document vector. Wordnet is represented by the data source titles *Auxiliary Source* in Figure 2. So for each term t in a document d if the term has a hypernym, we retrieved it and added it to the document representation. So our documents now contain each term and its hypernym $d =< t_1, t_2, \dots t_n, h_1, h_2, \dots h_k) >$. We then applied K-means to the document sets and recorded the results. The same procedure was performed on the synonym dataset. However rather than adding hypernyms, synonyms were instead included in the documents.

4.4 Graph Approach

For each of our main threads, we identified a thread dealing with a similar topic that would hopefully use the same, or similar, terminology. This is important for constructing a related ontology, e.g the thread *rugbyunion* is seen to be a related thread to *Australian Rugby League (NRL)*. Table 2 contains a list of the initial threads and the related threads from which we constructed our ontology. Our external ontology is represented by the data source titles *Auxiliary Source* in Figure 2.

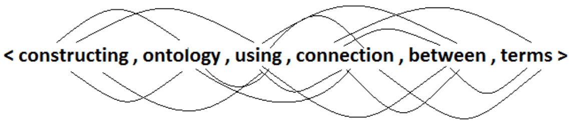
< constructing , ontology , using , connection , between , terms >

Figure 3: Recording the connection between terms.

To construct the ontology, we processed the related threads. As above, lemmatisation and stopword removal was applied to the document vectors. Next we constructed a graph where each word is represented as a node and the weight on the edge represents the number of times two words occurred in close proximity. We used a window size parameter to define the notion of proximity. In this work we use a window size of two, i.e., the nearest two preceding and proceeding words for each word are considered as occurring in close proximity and their corresponding nodes are linked. Figure 3 illustrates how two words can be connected. Our ontology comprises of the term t and all of the occurrences of the surrounding terms t_r and their frequency, $t = \{tr_1 : score, tr_2 : score, \dots tr_n : score_n\}$. Next, we augmented the vectors representing the original document by augmenting the document vector with its highest correlated word. The resulting documents were stored as follows: $d =< t_1, tr_1, t_2, tr_2 \dots t_n, tr_n >$. We applied the K-means clustering algorithm to the resulting corpus and recorded the results.

4.5 Process

Ten social media conversation threads were randomly selected from the social media platform Reddit. Threads were allocated into two groups of 5 threads called Thread A and Thread B (see Table 1). Threads from the first group were then paired with threads from the second group using a round-robin approach. This resulted in the creation of 25 document sets containing two separate threads each. In order to measure prediction performance, documents were subjected to four independent prediction processes: *Standard* which serves as the baseline (standard clustering), *Synonym* (document augmentation with synonyms), *Hypernym* (augmentation with Hypernyms), and finally *Graph*, our approach which augments the documents with the strongest correlates according to our graph measure over the bespoke ontology. Prediction performance was measured by the number of errors in identification of separate threads within a document. Lower levels indicated greater prediction performance.

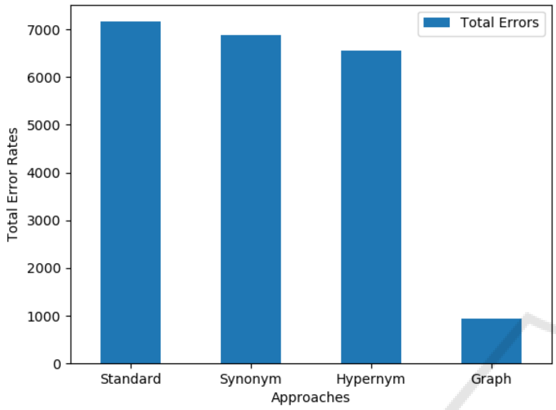

Figure 4: Combined error rate over all tests.

5 RESULTS

Table 3 presents the results from each individual clustering task. It contains the name of the two threads being investigated; the number of documents being clustered in that case, and the number of errors for each approach. We used precision as a metric for evaluating the success of each approach. We found that adding synonyms and hypernyms improves upon the baseline case. Echoing findings in previous work, we found that hypernyms are more effective than synonyms. Finally, we note that using our graph approach is six times more effective than using any of the other approaches. The combination of using a bespoke ontology and a graph mechanism to identify correlates for expansion works extremely well.

5.1 Statistical Analysis

A non-parametric Kruskil-Walis (Cohen, 1988) test was conducted to explore whether using a Graph analysis approach would improve prediction performance of the presence of separate threads in a body of text when compared with three commonly used approaches. This paper makes the hypothesis that using the graph approach will result in lower prediction errors. No hypothesis is made between prediction levels of the three commonly used techniques when compared to each other. Analysis was conducted using the computer software package SPSS.

Results indicate that Graph approach reported fewer errors in identification (M =4.72, SD = 9.14) to Approach Standard (M =35.97, SD = 38.48), Approach Synonym (M =34.58, SD = 36.90) and Approach Hypernym (M =32.91, SD = 35.94).

A Kuskal-Wallis test revealed a statistical difference in prediction performance across the four prediction approaches $H(3) = 95.19, p < .001$. Pairwise comparisons with adjusted p values showed there was a significant difference between the graph approach and the standard approach $(p < .001, r = .4)$ the synonym approach, $(p < .001, r = .4)$ the hypernym approach, $(p < .001, r = .4)$, indicating medium to strong effect sizes. The standard, synonym and hypernym approaches did not differ significantly from each other $(P > .05)$

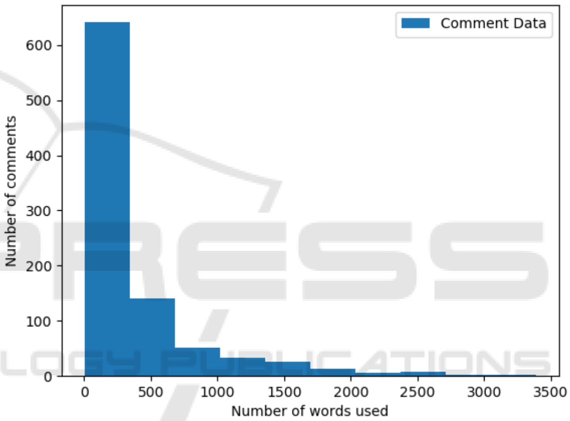

Figure 5: Break down of comment lengths.

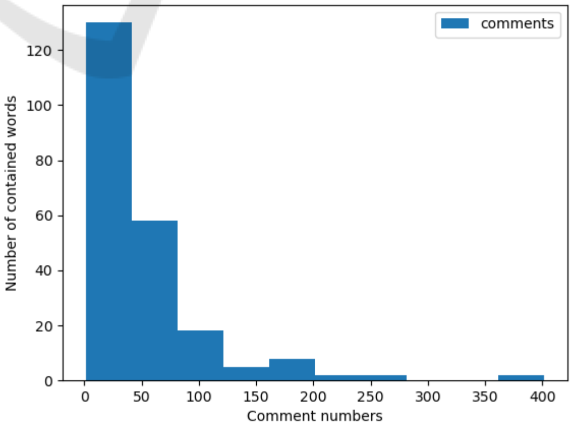

Figure 6: Break down of comment lengths from misclassified comments.

Table 3: Results for each algorithm or each combination of threads from the first and second group respectively.

index	Name	Size	Standard	Synonym	Hypernym	Graph
0	rugbyunion_quadcopter	1598.0	215.0	215.0	226.0	43.0
1	rugbyunion_Worldnews	1439.0	373.0	201.0	123.0	22.0
2	rugbyunion_politics	1889.0	372.0	438.0	414.0	1.0
3	rugbyunion_boardgames	1868.0	283.0	289.0	253.0	48.0
4	rugbyunion_Ireland	1861.0	337.0	329.0	333.0	31.0
5	learnpython_quadcopter	1576.0	105.0	100.0	109.0	23.0
6	learnpython_Worldnews	1417.0	178.0	124.0	114.0	4.0
7	learnpython_politics	1867.0	373.0	446.0	411.0	0.0
8	learnpython_boardgames	1846.0	220.0	219.0	201.0	19.0
9	learnpython_Ireland	1839.0	235.0	219.0	211.0	19.0
10	Movies_quadcopter	1094.0	126.0	122.0	132.0	39.0
11	Movies_Worldnews	935.0	376.0	131.0	127.0	23.0
12	Movies_politics	1385.0	454.0	527.0	515.0	20.0
13	Movies_boardgames	1364.0	258.0	262.0	231.0	21.0
14	Movies_Ireland	1357.0	153.0	151.0	150.0	32.0
15	music_quadcopter	1120.0	156.0	138.0	132.0	51.0
16	music_Worldnews	961.0	153.0	134.0	124.0	27.0
17	music_politics	1411.0	445.0	521.0	503.0	2.0
18	music_boardgames	1390.0	109.0	103.0	91.0	16.0
19	music_Ireland	1383.0	154.0	136.0	133.0	39.0
20	England_quadcopter	1415.0	320.0	251.0	219.0	84.0
21	England_Worldnews	1256.0	378.0	380.0	378.0	81.0
22	England_politics	1706.0	426.0	491.0	470.0	2.0
23	England_boardgames	1685.0	294.0	286.0	283.0	67.0
24	England_Ireland	1678.0	666.0	668.0	667.0	226.0

Table 4: Data Facts.

	Standard	Synonym	Hypernym	Graph
count	25.0	25.0	25.0	25.0
mean	286.36	275.24	262.0	37.6
std	136.05	159.65	157.88	45.52
min	105.0	100.0	91.0	0.0
max	666.0	668.0	667.0	226.0

6 DISCUSSION

To gain a better insight into why our approach achieves better results, we analysed some of the characteristics of the thread data and how the methods applied affected the clusters. In Figure 5, we show the thread *Rugbyunion* as an example; we first plotted the distribution of the sizes of each of the threads. From this it is clear to see that a large number of the comments are between 0 and 500 words in length. Figure 6 is a break down of the sizes of the comments that were misclassified. We can tell from this that the highest level of misclassification comes from documents that are 30 words or less in length. This makes intuitive sense when one considers that the less evidence the classifying agent has, the poorer the end result will be. Our graph approach helps to offset this issue, by incorporating words that are more indicative of the document class thus producing significantly more accurate classifications. The sum total of words added for each method were Synonym - 285,321; Hypernym - 220,935 and Graph 304, 997. Of these additional words the number of unique words added were 9435, 3584 and 4023 respectively. While the Hypernym had the least number of unique terms it also had markedly less terms added compared to the other two approaches. The Synonym approach had a large number of additional terms, although there were a little less than the Graph approach which had the most terms added, but a relatively low unique word count. This leads us to conclude that the terms returned were more closely correlated in this approach, which resulted in the higher precision counts.

Figure 7 offers some more insight into performance of the various algorithms across the different clustering cases. The graph approach is clearly superior to the other approaches with a much lower median number of errors but also a much smaller deviation. The orange line represents the median line, and interestingly it is higher in the standard approach. This means that in over half of the clusters, the hypernym and synonym approach are superior. The whiskers are higher in both of these approaches suggesting that there is a large variation in a small number of results. This suggests that while the addition of hypernym and synonym did, on average, improve the results, there are a minority of instances where they added noise to the dataset and skewed some of the results. This phenomenon is not witnessed in the graph approach which only improved upon results.

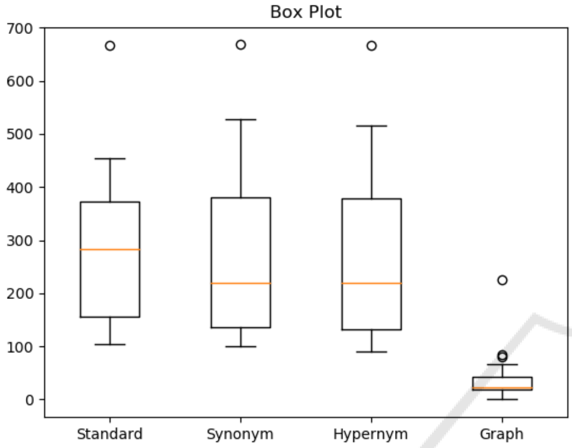

Figure 7: A box plot of all of the errors.

7 CONCLUSIONS

This paper discusses the investigation of the use of external ontologies to improve performance of a clustering algorithm through meaningful augmentation of documents. A standard package Wordnet was used to identify if the use of hypernyms and synonyms can improve performance. Additionally a bespoke ontology was constructed that represents relationships between terms based on co-occurrence, to see if the use of context can improve results. Our dataset is not a standard document collection so it poses additional challenges that limit the effectiveness of traditional clustering approaches. The best results were shown to be achieved when context was used.

In this work, we manually identified related threads from which to construct our ontology. Future work will see the automatic identification of related reddit threads. This can be achieved through measuring syntactical similarity between threads. The future aim of our work is to enhance this context construction and applying it to identifying different points of view.

REFERENCES

Baghel, R. and Dhir, R. (2010). A frequent concepts based document clustering algorithm. *International Journal of Computer Applications*, 4(5):6–12.

Bennett, W. L. (2012). The personalization of politics political identity, social media, and changing patterns of participation. *The ANNALS of the American Academy of Political and Social Science*, 644(1):20–39.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2002). Latent dirichlet allocation. *Advances in neural information processing systems*, 1:601–608.

Bruza, P. and Song, D. (2001). Discovering information flow using a high dimensional conceptual space. In *Proceedings of the 24th ACM SIGIR Conference on Research and Development in Information Retrieval*.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391.

Golkar, S. (2011). Liberation or suppression technologies? the internet, the green movement and the regime in Iran. *International Journal of Emerging Technologies and Society*, 9(1):50.

Hotho, A., Staab, S., and Stumme, G. (2003). Ontologies improve text document clustering. In *Data Mining, 2003. ICDM 2003. Third IEEE International Conference on*, pages 541–544. IEEE.

Hung, C., Wermter, S., and Smith, P. (2004). Hybrid neural document clustering using guided self-organization and wordnet. *IEEE Intelligent Systems*, 19(2):68–77.

Lotan, G., Graeff, E., Ananny, M., Gaffney, D., Pearce, I., et al. (2011). The arab spring— the revolutions were tweeted: Information flows during the 2011 tunisian and egyptian revolutions. *International journal of communication*, 5:31.

Lund, K. and Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, & Computers*, 28(2):203–208.

Mahajan, S. and Shah, N. (2016). Efficient pre-processing for enhanced semantics based distributed document clustering. In *Computing for Sustainable Global Development (INDIACom), 2016 3rd International Conference on*, pages 338–343. IEEE.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Shah, N. and Mahajan, S. (2012). Document clustering: a detailed review. *International Journal of Applied Information Systems*, 4(5):30–38.

Shirky, C. (2011). The political power of social media: Technology, the public sphere, and political change. *Foreign affairs*, pages 28–41.

Singer, P., Flöck, F., Meinhart, C., Zeitfogel, E., and Strohmaier, M. (2014). Evolution of reddit: from the front page of the internet to a self-referential community? In *Proceedings of the 23rd International Conference on World Wide Web*, pages 517–522. ACM.

Wang, Y. and Hodges, J. (2006). Document clustering with semantic analysis. In *System Sciences, 2006. HICSS'06. Proceedings of the 39th Annual Hawaii International Conference on*, volume 3, pages 54c–54c. IEEE.

Weninger, T., Zhu, X. A., and Han, J. (2013). An exploration of discussion threads in social news sites: A case study of the reddit community. In *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pages 579–583. ACM.

Zheng, H.-T., Kang, B.-Y., and Kim, H.-G. (2009). Exploiting noun phrases and semantic relationships for text document clustering. *Information Sciences*, 179(13):2249–2262.