Visual Analytic System for Subject Matter Expert Document Tagging using Information Retrieval and Semi-Supervised Machine Learning

Craig Hagerman
Uncharted Software
Toronto, Canada
chagerman@uncharted.software

Richard Brath
Uncharted Software
Toronto, Canada
0000-0001-6006-2092

Scott Langevin
Uncharted Software
Victoria, Canada
slangevin@uncharted.software

***Abstract*—We present a system that combines ambient visualization, information retrieval and machine learning to facilitate the ease and quality of document classification by subject matter experts for the purpose of organizing documents by "tags" inferred by the resultant classifiers. This system includes data collection, a language model, query exploration, feature selection, semi-supervised machine learning and a visual analytic workflow enabling non-data scientists to rapidly define, verify, and refine high-quality document classifiers.**

Keywords—visual analytics, classifiers, machine learning, ambient visualization, mixed-initiative analytics, information retrieval.

I. Introduction

Ambient visualizations convey time-varying information in the periphery of human attention [1]. As these visualizations need to maintain interest without interaction, data must be in a form to facilitate updates. In this research project, we extended a large-scale ambient visualization system beyond its original use for quantitative data into unstructured data such as news, research documents, and other documents. A key challenge involved how to classify large corpora of unstructured data by topics of interest in order to categorize by tagging the data into relevant subsets of high quality, appropriately classified items displayed using ambient visualizations.

The research was motivated by off-the-shelf data sources that were either unlabelled (i.e. not classified) or insufficiently labelled; labelled data either did not correspond with topics of interest to subject matter experts (SMEs) or contained false positives, false negatives or otherwise mislabelled items.

This research sought to develop a methodology and system to (1) provide user-in-the loop labelling of data feeds of unstructured data, (2) provide rapid SME-driven classifier construction leveraging information retrieval (IR) and machine learning (ML) classification to automate data labelling, and (3) a visual analytic workflow system to construct, view, validate, and interact with derived classifiers.

This work was supported, in part, by the Defense Advanced Research Projects Agency (DARPA) (contract number MEMEX (FA8750-14-C-0275)). The views, opinions, and findings contained in this report are those of the authors and should not be construed as an official Department of Defense position, policy, or decision.

Visualization, IR and ML classification approaches are mature areas of research. Ambient visualizations are not art; they must be decodable [2,3], and need to attract [4]. Traditional IR focuses on ranking documents by relevance to a query [5]. Rather than manual trial-and-error query formulation, supervised ML techniques such as text classifiers automatically categorize documents into classes of interest based on labelled training data consisting of ground truth class membership. One distinction between IR and supervised ML approaches is the ML requirement for labelled data to determine historical class membership. The advantage of ML augmentation of IR approaches is the automatic categorization, ranking, and filtering of documents into classes of interest given past examples versus laborious manual trial-and-error keyword queries for relevant data. A challenge with constructing ML classifiers is the high expense in time and resources needed to label training data to construct a high-performing classifier. Semi-supervised ML aims to reduce this expense by labelling a smaller amount of data and bootstrapping the learning process by using unsupervised ML to infer labels of the unlabelled training data to construct a larger labelled training dataset that constructs the classifier. The accuracy of the resultant model relies heavily on the quality of the candidate dataset that is manually labelled to bootstrap the process.

In this paper we present a method of combining IR and ML classification in an iterative visual analytic workflow to enable an SME to rapidly construct high-quality domain-specific document classifiers with minimal effort and no data science expertise. The SME should be able to validate the performance of the resulting classifier using novel ambient visualizations. Our approach aims to drastically reduce the time required for an SME to label a large corpus, while ensuring the training data is representative of the classification task to produce an accurate classifier. A human-in-the-loop process uses semi-supervised learning to automate generation of the classifier based on these annotations.

II. Background

A. Text Classifier Challenges

Supervised classification techniques require labelled data. For example, in a binary classification, examples are needed for both the positive and negative classes of interest. This labelled data can then be used as training data for the classifier. This is

typically a time-consuming job for human annotators, requiring SMEs with domain expertise to label a large corpus of data in order to construct a reliable classifier. Annotating large datasets can be prohibitively expensive in terms of the time required to label data and in the cost of paying an SME for the task. There are four sub-problems that make human annotation of a dataset a challenge: (1) time to annotate an individual document is too high; (2) there are too many documents to annotate individually; (3) utilizing existing annotation systems sometimes involves specialized data science skills [6]; and (4) an easy-to-use interface is needed to streamline the above tasks.

Labelling data is foundational for any supervised learning task. Identifying good candidates to label is important to improve a classifier's performance [7].

With a sufficiently large corpus of labelled data, a classifier can be trained to produce a statistical model that will generalize and provide accurate predictions on new data. However, even if significant effort is spent labelling data, it is not always obvious that the training data is representative of the distributions in the underlying domain of interest.

B. Reducing Barriers to Text Classification

A primitive document annotation task might provide SMEs with raw text files alongside a spreadsheet where documents are associated with class labels. SMEs would be tasked with reading a text file and then switching to the spreadsheet to insert a class label associated with that document. While such an annotation spreadsheet is a useful form for machine learning tools, the workflow and time required for the user to annotate a large document set quickly becomes tedious and error prone. Thus, a tool that can streamline the process would benefit SMEs.

In addition to user interface design goals that streamline the annotation process of individual documents, reducing the overall number of documents that require annotation is also an approach to reduce the overall cost of annotation. One approach is building an interactive, Active Learning system [8]. Another, for a binary classifier, is to eliminate the need for negative labelled examples. Li and Liu [9] describe a semi-supervised learning approach that requires only positive examples be labelled to construct a binary classifier.

We propose to improve document classification using a combination of several techniques: (1) the use of query expansion to recommend potential positive examples to label using alternate spellings, and a word embedding model for semantic similarity of terms and phrases; (2) an iterative process that builds the classifier and allows the SME to preview results; (3) using these interactively defined text classifiers to search document repositories for positively matched documents, for review and validation by SMEs; and (4) using visualizations to aid assessment and refinement of the classifier.

C. Visualization of Classified Text

The target use of the classified results is an ambient visualization system. This system has six animated visualizations. Each visualization provides an initial overview of the dataset, animates to specific observations, such as a callout of the largest values, and simple interactions to further explore the data. This follows the martini glass structure of narrative visualization [10]. For example, a 2D world map depicts geospatial data followed by animated call-outs to indicate the largest values and then the ability to tap to show any specific values. A scatterplot animates data points over time, then animates callouts of particular values such as outliers, and provides filtering of categories. While each of these visualizations was previously designed for quantitative data, the primary goal is to reuse them for animated display of the textual data, with a secondary goal to use them to facilitate the text classification process.

III. Technical Approach

Our proposed approach is to annotate (i.e. label) a corpus through an iterative process that builds text classifiers based on human guidance. This approach combines information retrieval with semi-supervised learning to allow SMEs to rapidly prototype, refine and train high-quality classifier models using an interactive workflow that includes visualization.

The SME starts with a text search query for particularly interesting or relevant documents using keywords and phrases known to be of domain relevance. In the query exploration step, a query expansion engine proposes semantically similar and related terms, while also retrieving and displaying sample documents that match the nominated terms. As part of a feature selection step, the SME can then indicate the relevancy of particular phrases and terms – both those they entered and those proposed by the similarity service.

After the SME has chosen a set of query terms and phrases, the information retrieval system identifies and labels all returned documents as positive examples. A semi-supervised classifier then uses this set of positive examples to train a classifier to identify more documents in the unlabelled data that should be positive. The visualization can be used to assess the results of the classification and guide further iterations.

A. Data Sources

In our approach, we first collect data from a variety of sources. For a labelled dataset for validating our approach, we use the canonical Reuters-21578 text collection [11]. An applied domain of interest for our research is rapid categorization of news events. For representative large event corpora, we use two sources: one a commercial news data source, the other news events extracted from GDELT [12], an open-source data service of machine extracted events (with a source, title, event type, actors, dates, geo, and other metadata attributes) from world news sources. We used GDELT to extract the top *n* events and fetch the associated article metadata and lead paragraph. Over the course of a year, we collected tens of thousands of news events.

B. Language Model

The similarity service uses language models to identify semantically similar and related terms and documents. For word and phrase level similarity, we combine two approaches: word embeddings similarity derived from a Word2vec model [13] and Elasticsearch edit distance similarity [14]. The Word2vec software library can generate word embeddings or a model of a word vector space.

For document-level similarity, we use three different text vectorization approaches: term frequency – inverse document

frequency (TF-IDF), averaging the word embedding vectors of each word in a document, and Doc2vec [15]. Doc2vec is analogous to word embeddings, but represents documents rather than words as a vector.

C. Query Exploration

The Query Exploration phase helps an SME to identify keywords and phrases most relevant to their query. Although an SME will understand terms relevant to their search, semantic similarity services aid the SME by offering phrases and keywords they had not considered, including misspellings and alternate spellings. The similarity service also allows the SME to quickly narrow and define their search criteria by identifying and excluding connotations not relevant to their search.

The goal of the query exploration component is to: (1) improve discoverability within the corpus, (2) reduce cognitive load on the SME by helping identify similar or potentially relevant terms, (3) interactively identify related terms that are relevant; and (4) quickly partition the data.

The similarity service is a REST server that takes an SME query and requests: (1) alternate and common misspellings of terms and phrases from Elasticsearch, using the built-in completion suggester [13]; and (2) related terms and phrases from Word2vec word embedding models [14]. To incorporate domain-specific vocabulary into the semantic similarity service, Word2vec models are trained for both terms and phrases using a representative document corpus. If a query term is not part of the domain vocabulary, the service falls back to a pre-trained Google News word embeddings model to incorporate a generic vocabulary.

In practice, the SME receives various auto-completion suggestions, semantically similar suggestions, and common variations of their terms as they are typing in a query. The results appear as easily deletable tags, such as *gas* or *gulf* (Fig. 1).

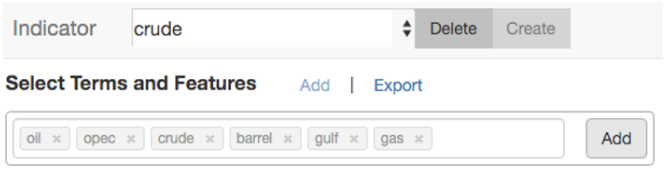

Fig. 1. Interface with classifier name (top) and set of expanded query candidates (bottom).

D. Feature Selection

The feature selection interface allows the SME to make quick judgements about the relevance of large numbers of documents with minimal interactions. The application presents, within the context of their containing documents, string matches for the exact query the SME provided and for expansions identified by the query exploration service.

For each term or phrase, the IR portion of the system retrieves high-ranking documents that match the terms and phrases. Snippets of these documents are shown with the expanded search terms in context and highlighted so the SME can understand how the terms and phrases are used in context of the corpus (i.e. keywords in context (KWIC) [16]).

After reviewing these snippets from the actual corpus, it is possible that some of the terms or phrases are not analytically useful as indicators for the class of interest. Alternatively, it is possible that a specific term entered by an SME is not relevant to a domain classifier when seen in the context of example documents. In this case, the SME can mark a term or phrase as irrelevant so that it is excluded when building the classifier. This allows SMEs to check whether specific terms or phrases would be relevant to a domain classifier and quickly exclude them if they are not.

Fig. 2 shows a snapshot for indicator *crude* and additional seed terms from the query expansion. Note that *barred* and *barren* were nominated as alternate spellings for the seed *barrel*, which incorrectly leads to irrelevant documents. The SME can select the entry for *barrel*, preview the resulting documents and turn off the incorrect variations.

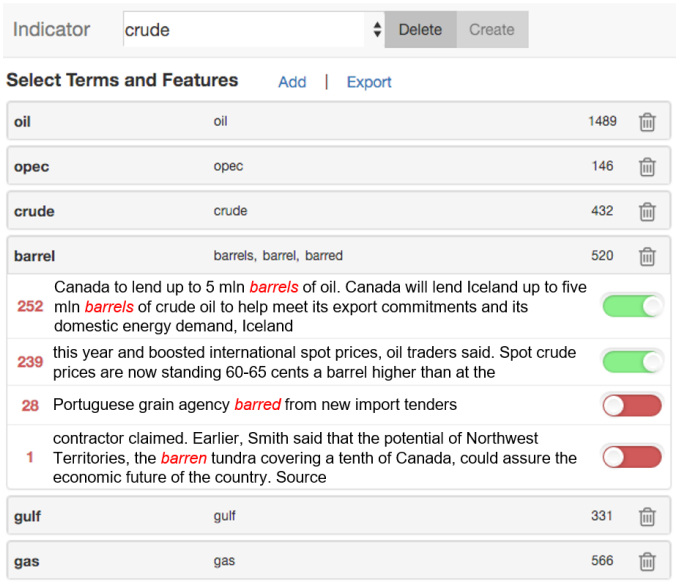

Fig. 2. Interface showing keywords highlighted in context for high-ranking documents, with explict switches to label the associated documents as positive or negative.

An SME marking a term or phrase as relevant is equivalent to selecting the documents that match as positive training examples for the class of interest. Documents in the corpus that do not match the list of terms/phrases and documents that are explicitly excluded as not relevant act as “unlabelled” examples. Using a small set of query terms, an SME can quickly label tens of thousands of examples as relevant. Following this labelling step, the matching positive examples are used in a semi-supervised learning process to construct a text classifier.

Furthermore, each document can be selected to see examples of similar documents using a Doc2vec model for document similarity. Each document returned by the IR system (and not toggled off) is labelled as a positive class. Each is initially assumed to be an exemplar of the positive class, and document similarity is used to surface additional documents to include in the positive class. In the same way, documents toggled off are assumed to be exemplars of the negative class and Doc2vec is used to discover and label further documents. Approximate

Nearest Neighbors (ANN) [17] is used to quickly find similar documents.

This semi-automated query expansion and feature selection allows an SME to quickly explore and partition a very large corpus.

E. Semi-supervised learning

Our semi-supervised classifier, based upon Li and Liu's approach [9], learns from positive and unlabelled examples. For initial development we used Python to train classifiers using Scikit-learn, before reimplementing a scalable distributed version in Apache Spark. This implementation allows us to scale the approach to larger datasets by distributing workload across a compute cluster. We also utilize the MLlib Spark library, which has implementations for common text preprocessing algorithms (such as stop word removal), document vectorization approaches, and common machine learning algorithms [18].

The approach outlined extracts reliable negative documents from the unlabelled set and builds an effective classifier for the positive class. The positive labelled documents in our proposed technique are curated from the user interface described in previous sections. and are passed to the system outlined in Fig. 3.

In the absence of true negative examples, the Li and Liu method first considers all unlabelled documents U as negative N and then uses the positive examples P and unlabelled U as the training data to build a binary classifier. The classifier is then used to classify U into a new positive set P' and a reliable negative (RN) set.

During the feature selection phase in our implementation, the SME may nominate documents as non-relevant. These, along with Doc2vec similar documents, are removed from P and combined with U for classification.

The Li and Liu method only uses the TF-IDF vectorization scheme. Our results in the Evaluation section show that the Word2vec and Doc2vec vectorization approaches improve performance in the classification models. Additionally, our query exploration and feature selection approach allows an SME to rapidly create a high-quality positive example set.

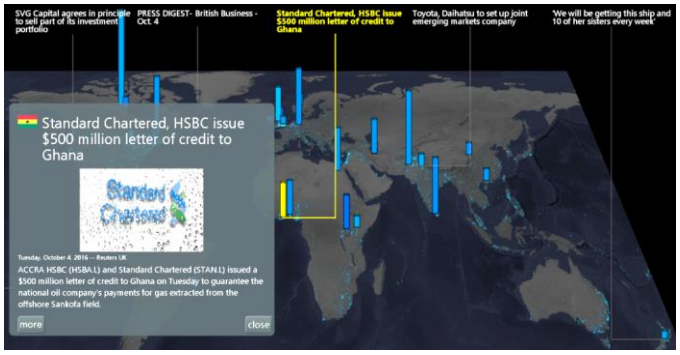

Fig. 3. Map view showing all news as dots on map, with top stories shown as bars. Five headlines across the top are connected to bars by leader lines. Animated pop-up shows story details.

F. Visualization

Our system has two visualizations relevant for the news domain to show results from a trained classifier. A map visualization geographically plots all matching documents for a given classifier as thousands of dots on a base map, with 10-20 widely reported stories shown as bars (Fig. 3). As an ambient display, the map first animates the most relevant headlines (with corresponding leader lines to the geographic location of the content) across the top of the screen. Next, an animation progresses through each headline, displaying the corresponding lead paragraph, any photos, and other relevant content.

The system also contains a scatterplot which can show dots, labels or full sentences such as headlines. These are color coded to different classifiers (e.g. oil, corn, metals). There are alternatives for the configuration of the coordinate space to layout the data points: 1) explicit x- and y-axes such as recency (x), number of sources (y), with an optional depth (z) axis (Fig. 4); 2) spatial coordinates can be derived by placing the documents in a high-dimensional vector space (e.g. Doc2vec) and reducing down to 2 or 3 dimensions using principal component analysis (PCA) or another dimensional reduction approach. 3) randomized layout, like a word cloud, animating different classes from foreground to background along the depth axis (Fig. 5).

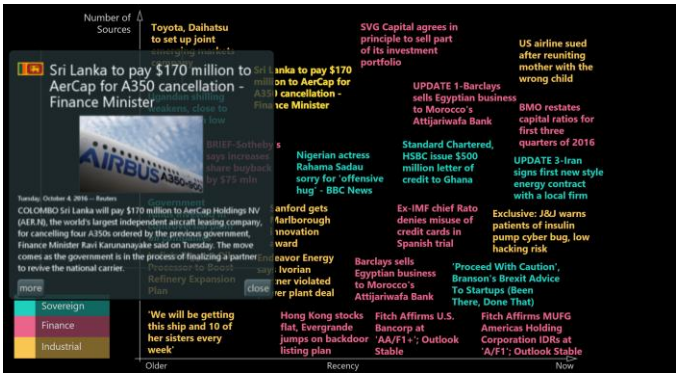

Fig. 4. Scatterplot configured with x-axis indicating recency and y-axis indicating number of sources.

These visualizations provide valuable feedback on the quality of the result set. For example, the classifier for *corn* has many documents related to corn-producing or corn-consuming regions of the world. The SME previewing documents can use preexisting knowledge to confirm the quality of the results. This combination of visualization content preview and visualization helps SMEs build and validate classifiers and obviates the need for machine learning experts.

Fig. 5. Scatterplot as a word cloud of headlines. Headlines animate in depth so that a particular class comes forward.

IV. Evaluation and Discussion

To evaluate the proposed approach, we use multiple evaluation mechanisms: (1) verification that the semi-supervised learning approach performs comparably to prior work; (2) end-user feedback that, together with the elements of the system, provided an effective solution to classifying and presenting content to users; (3) comparison between the documents classified by a user using our approach versus ground truth labels present in a small sample of commercial news event data; and (4) feedback from SMEs on the important terms and phrases surfaced both by the iterative query exploration phase and those identified by the classifier as highest relevance.

A. Semi-supervised learning.

We applied the system to the canonical *Reuters-21578* text collection [11] in the same manner as discussed in Li and Liu. The top ten categories (by number of articles) are chosen from the Reuters dataset. This defines ten positive and unlabelled classification problems. Query exploration was used to discover terms and phrases representing the classes. Features were selected and documents retrieved for those features. The semi-supervised learning system was run on the classes; here, the positive set is the documents from the class of interest and the unlabelled set is the documents in the remaining nine categories.

TABLE I. F1 Score Results

	SVM Baseline	Li Results	Spark Impl	Improved
acq	0.94	0.905	0.939	0.898
corn	0.9	0.635	0.611	0.753
crude	0.89	0.811	0.89	0.871
earn	0.98	0.886	0.865	0.893
grain	0.95	0.903	0.911	0.923
interesr	0.78	0.614	0.68	0.769
moneyfx	0.75	0.764	0.777	0.776
ship	0.86	0.829	0.774	0.848
trade	0.76	0.728	0.714	0.775
wheat	0.92	0.779	0.738	0.764
	0.873	0.7854	0.7899	0.827

We present results for a baseline supervised classifier, our implementation of the algorithm in Li and Liu, and an improved version of the Li and Liu algorithm. Table 1 shows F1 scores for each variation. The F1 score, or F measure, is a common measurement of binary classifier performance that equally weights precision and recall scores.

The SVM baseline column shows the results of a supervised classification approach for comparison with the semi-supervised approaches used in this research. The baseline performs excellently in all ten categories, with an average F1 score of 0.873. The original results from the Li and Liu paper perform worse than the baseline, with an average F1 score of 0.7854. With our Spark implementation of the Li and Liu algorithm, we achieved an average F1 score within 0.57% of the performance of the original paper. The last column presents the results for our improved version of the Li and Liu approach. We made two key improvements: (1) we replaced the TF-IDF vectorization scheme in the original paper with a word embedding approach; and (2) we replaced the SVM approach used to build the final classifier with a logistic regression model. These resulted in an average F1 score of 0.827, which is only 5% worse than the supervised learning baseline, and better than the results from Li and Liu's approach.

B. Comparison with another system

We compared our results to a a large, legacy, manually curated rule-based classification system to annotate data. Based on the data available to us from the rule-based system, accuracy was similar between approaches. The ability to interactively define machine learning classifiers rather than manually construct and test rules, and the ease of maintaining and refining the classifier were seen as notable differences. User feedback is described in the next section.

Some of the differences in labelling were revealing. Closer examination of some conflicting tags revealed false positive and false negatives in the rule-based system seemed to be the result of manually curated binary classification rules. For example, a story about a baseball team called the *Corn-Belters* was labelled as belonging to the classification for *corn*, due to the presence of an important corn keyword.

C. End user feedback.

The prototype system was reviewed with different user groups. Management is responsible the conceptualization, funding and roll-out of the ambient visualization across locations. SMEs are responsible for the content. Various issues were raised:

- *Set and forget.* While management liked that new classifiers for emerging topics could be easily created and updated (e.g. *Brexit*), users were more interested in setting up stable topics in a one-time configuration rather than creating new topics. While stakeholders would generally ask for configurable functionality, in practice they would rarely modify the configuration.
- *Relevance.* Some stories are more relevant than others. A story about corn producers in Iowa impacted by trade policy is much more relevant to SMEs than stories about corn at a summer picnic. We implemented a scoring system for different sources to differentiate between relevancies.
- *Interestingness.* Correctly classified, highly relevant documents may still be uninteresting. A crop report with no unexpected information is not interesting. This was not addressed.
- *Blacklist by tone.* Negative stories regarding the organization or their customers were not desired. We implemented a configurable blacklist to filter out stories corresponding to specific keywords.
- *Map.* Users liked the map-based representation, which provided global orientation and context. In addition to confirming news relevance, the visualization also aids ambient consumption. For example, if the topic *corn* was the focus, the map effectively orients expert users to

stories about different participants, such as producers (e.g. USA, Brazil) or importers (e.g. Japan, Mexico). It also helps SMEs building the models to match data to real-world expectations and visually detect anomalies.

- *Scatterplot.* User reactions to the scatterplot were mixed. In general, users preferred an immediately understandable representation (i.e. very low cognitive effort). Multidimensional reduction was too difficult to explain, comprehend, and was dismissed. We believe that multidimensional reduction visualization will be beneficial to SMEs building the corpus, but this has not been implemented yet.

  A few users liked the explicit axes such as recency and sources. They understood the logical explanation, but most were otherwise unenthusiastic. We hypothesize that mapping data points back to axes and quantities presumably requires extra effort to create a mental model of the cartesian space. Only with familiarity and practice would such a visualization achieve low cognitive effort.

  Finally, the word cloud variant was also unenthusiastically received. The stakeholders previously had some missteps with ambient visualization that were not decodable – in effect *information art* instead of *information visualization* [19]. The organization has a high proportion of analytical staff that want usable information. The word cloud variant was perceived as low informational content.

- *Source quality and access*: An ongoing issue was the quality and quantity of data feeds. Open source feeds (e.g. GDELT) were broad but with abundant low-quality news and irrelevant stories. High-quality sources were expensive and difficult to access. Scraping data was briefly considered but abandoned due to an increased level of effort.

D. Discussion of Results

The user feedback suggests that end-user perception of the value surrounding classification and visualization of textual documents is broader than classification and presentation. Relative relevance, interestingness based on unknown information, and exclusion by tone are filters beyond the scope of requirements. Despite working with SMEs and end-users throughout the process, these only became apparent with real-world data sufficiently well classified in order for these issues to appear.

The disinterest in all variants of scatterplots was unexpected, presumably due to understanding the user community's acceptance of abstraction and ease of decoding. In the nested model of visualization design and validation [20], this is an error at the abstraction level. This mismatch does not align with definitions of ambient visualizations that might prioritize aesthetics over data [19, 21]. This mismatch was hinted at by the users' prior issues with earlier ambient visualizations.

We experimented with two approaches to clustering documents to facilitate the iterative query exploration - feature selection phase: nearest neighbor search and topic modeling.

We used Doc2vec vectors with ANN search to identify other documents within the neighboring vector space near nominated documents. Although this approach showed promise, it did not improve classification on the limited labelled data from the other system. We hope that in future work with larger datasets, the benefits of ANN-aided partitioning of data will become clear.

We also experimented with Latent Dirichlet Allocation (LDA) topic modeling to (1) identify topic clusters of documents and (2) identify the most salient keywords for each, possibly identifying new tags of which the SME was unaware. This approach identified clusters in the pre-labelled data, and thereby aided partitioning. However, this approach was not included in the prototype since a non-parametric topic labelling approach (i.e. where the number of clusters is not known in advance) would be needed for unlabelled, unstructured data.

V. Conclusions

We created a system for the collection, human-in-the-loop classification and data visualization of unstructured data such as news. We demonstrate a workable approach where the combination of IR and ML technologies and a map-based visualization is greater than the use of each in isolation. We show an approach for SMEs to easily label the data and validate the approach. We show issues broader than the immediate scope which were revealed only after a prototype: a desire for more metadata beyond classification such as relevance, interestingness and tone; and visual representations that are both aesthetically pleasing for ambient visualization and easily decodable.

Future work should include extending the scatterplot visualizations to aid the review of the documents during the labelling stage. Measuring interestingness is a future task, potentially leveraging counts of user interactions to label documents of highest interest.

We have found indications that additional approaches to partitioning the data during the feature selection stage, such as ANN search and non-parametric topic modeling, could lead to greater human-guided classification and tagging efficiencies. Given a corpus of tagged documents we could improve automated annotation by constructing a multi-tag classifier. A related avenue of future work could include incorporating a tag recommendation algorithm derived from a term-document bipartite graph [22,23] or ontological guidance [24,25].

References

[1] A. Moere, "Towards designing persuasive ambient visualization," *Issues in the Design & Evaluation of Ambient Information Systems Workshop*, Citeseer, 2007.

[2] A.Lang, "Aesthetics in Information Visualization," *Issues in the Design & Evaluation of Ambient Information Systems Workshop*, Citeseer, 2007.

[3] R. Kosara, "Visualization criticism-the missing link between information visualization and art," In 2007 11th International Conference Information Visualization (IV'07), pp. 631-636. IEEE, 2007.

[4] B. Bafadikanya, "Attractive Visualization," In *Trends in Information Visualization*. D. Baur, M. Sedlmair, R. Wimmer, Y. Chen, S. Streng, S. Boring, A. De Luca, A. Butz, Eds. Technical Report LMU-MI-2010-1, Apr. 2010ISSN 1862-5207. University of Munich, Department of Computer Science, Media Informatics Group. 2010.

[5] C. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*, Cambridge University Press. 2009.

[6] J Pustejovsky and A Stubbs, 2012. *Natural Language Annotation for Machine Learning*, O'Reilly Media.

[7] J. Pujara, B. London, and L. Getoor, "Reducing Label Cost by Combining Feature Labels and Crowdsourcing," In Proceedings of the 28th International Conference on Machine Learning, 2011

[8] B. Settles, "Active learning literature survey," University of Wisconsin, Madison, 2010.

[9] X. Li and B. Liu, "Learning to classify texts using positive and unlabeled data," In Proceedings ofthe 18th International Joint Conference on Artificial Intelligence, 2003.

[10] E. Segel and J. Heer, "Narrative visualization: Telling stories with data." IEEE transactions on visualization and computer graphics 16.6: 1139-1148, 2010

[11] UCI Machine Learning Repository: Reuters-21578 Text Categorization Collection Data Set. archive.ics.uci.edu, 2016. https://archive.ics.uci.edu/ml/datasets/Reuters-21578+Text+Categorization+Collection.

[12] K. Leetaru and P. Schrodt, "Gdelt: Global data on events, location, and tone, 1979–2012." ISA annual convention. Vol. 2. No. 4. Citeseer, 2013.

[13] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient Estimation of Word Representations in Vector Space," arXiv:1301.3781

[14] Completion Suggester Elasticsearch Reference. Elastic.co, https://www.elastic.co/guide/en/elasticsearch/reference/2.1/search-suggesters-completion.html, 2016.

[15] A. Dai, C. Olah, and Q. Le. "Document embedding with paragraph vectors," arXiv preprint arXiv:1507.07998 (2015).

[16] M. Hearst, *Search user interfaces*, Cambridge University Press, 2009.

[17] G. Avarikioti, I. Emiris, I. Psarros, and G. Samaras. "Practical linear-space Approximate Near Neighbors in high dimension," arXiv preprint arXiv:1612.07405 (2016).

[18] X. Meng, J. Bradley, B. Yavuz, E. Sparks, S. Venkataraman, D. Liu, J. Freeman, D. Tsai, M. Amde, S. Owen, D. Xin, R. Xin, Michael J. Franklin, R. Zadeh, M. Zaharia, and A. Talwalkar. "Mllib: Machine learning in apache spark." The Journal of Machine Learning Research 17.1: 1235-1241, 2016.

[19] A. Lau and A. Vande Moere, "Towards a model of information aesthetics in information visualization," 2007 11th International Conference Information Visualization (IV'07). IEEE, 2007.

[20] T. Munzner. "A nested model for visualization design and validation," IEEE transactions on visualization and computer graphics 15.6: 921-928, 2009

[21] Z. Pousman, J Stasko, and M. Mateas. "Casual information visualization: Depictions of data in everyday life," IEEE transactions on visualization and computer graphics 13.6: 1145-1152, 2007

[22] Y. Song, L. Zhang, and C. Giles, "Automatic tag recommendation algorithms for social recommender systems,." ACM Transactions on the Web (TWEB) 5.1, 2011

[23] Z. Guan, C. Wang, J. Bu, C. Chen, K. Yang, D. Cai, and X. He, "Document recommendation in social tagging services," In Proceedings of the 19th international conference on World wide web (WWW '10). ACM, New York, NY, USA, 391-400, 2010.

[24] V. Ha-Thuc, Y. Mejova, C. Harris, and P. Srinivasan, "News event modeling and tracking in the social web with ontological guidance," IEEE Fourth International Conference on Semantic Computing, 414-419, 2010.

[25] V. Ha-Thuc, J. Renders, "Large-scale hierarchical text classification without labelled data," Proceedings of the fourth ACM international conference on Web search and data mining, 2011