Fast Text Classification Using Randomized Explicit Semantic Analysis

Aibek Musaev, De Wang, Saajan Shridhar, Calton Pu
Georgia Institute of Technology, Atlanta, Georgia
{*aibek.musaev, wang6, saajan, calton.pu*}*@gatech.edu*

***Abstract*—Document classification or document categorization is one of the most studied areas in computer science due to its importance. The problem is to assign a document using its text to one or more classes or categories from a predefined set. We propose a new approach for fast text classification using randomized explicit semantic analysis (RS-ESA). It is based on a state of the art approach for word sense disambiguation based on Wikipedia, the largest encyclopedia in existence. Our method reduces Wikipedia repository using a random sample approach resulting in a throughput, which is an order of magnitude faster than the original explicit semantic analysis. RS-ESA approach has been implemented as part of the LITMUS project due to a need in classifying data from Social Media into relevant and irrelevant items with respect to landslide as a natural disaster. We demonstrate that our approach achieves 96% precision when classifying Social Media landslide data collected in December 2014. We also demonstrate the genericity of the proposed approach by using it for separating factual texts from fictional based on Wikipedia articles and fan fiction stories, where we achieve 97% in precision.**

***Keywords*-text classification; explicit semantic analysis; social media; event detection**

I. Introduction

Automated document classification or document categorization is an important area in computer science. The problem is to assign a document using its text to one or more classes or categories from a predefined set. This technique is used in various domains, e.g. for detection of disasters like earthquakes [1]. The performance of text classification while maintaining high precision is especially important in case of real-time systems [2].

Our current area of interest is detection of landslides using an integration of multiple sources, including physical sensors and social networks like Twitter, Instagram and YouTube [3], [4], [5]. We use landslide related keywords, e.g. *landslide* and *mudslide*, to download items from social networks as input to our system. The challenge here is that they are polysemous words where one of their meanings is related to our domain and all other meanings are unrelated and introduce noise, including:

- *landslide* as an adjective describing an overwhelming majority of votes or victory: "Japan PM Abe's LDP on track for landslide in December 14 vote - media - World — The Star Online http://t.co/FrTbhnIazw"
- *landslide* as the Fleetwood Mac song "Landslide" from the 1975 album *Fleetwood Mac*: "Well I've been afraid of changing cause I built my life around you #LandSlide"
- *mudslide* as a popular cocktail: "The best dessert I found at Brightspot yesterday, not too sweet! @creamy-comfort #baileys #dessert #mudslide #brightspot brightspot"

A state-of-the-art approach in resolving the sense of polysemous words is called Explicit Semantic Analysis (ESA) and it was introduced by Gabrilovich et al. [6]. Their method represents the meaning of a text in a high-dimensional space of concepts derived from Wikipedia, the largest encyclopedia in existence. This approach, however, cannot be used for classification of texts directly due to the high number of dimensions, which is equal to the number of articles in Wikipedia. We propose to use a sample of the Wikipedia dataset instead of the full repository. This allows us to perform classification rapidly without necessarily having to make a large external repository of knowledge tractable first, while leveraging the capabilities of ESA as a superior word sense disambiguator.

This paper makes the following contributions:

- we introduce a generic approach for fast text classification using randomized explicit semantic analysis based on a random sample of Wikipedia articles (RS-ESA);
- we perform a quantitative evaluation of the proposed RS-ESA approach using real world landslide data collected in December 2014;
- we provide the results of comparison between the RS-ESA approach and the Expert-ESA approach where instead of a random sample of Wikipedia articles we use a set of related articles selected by an expert driven approach;
- we demonstrate the genericity of the proposed approach by successfully applying it to a different problem where factual texts are separated from fictional based on Wikipedia articles and fan fiction stories.

The rest of the paper is organized as follows. We describe the details of the proposed generic classification approach in Section II followed by the description of the expert based classification approach in Section III. We provide implementation notes in Section IV. In Section V we introduce all datasets that are used for experimental evaluation in Section VI. We summarize related work in Section VII and conclude the paper in Section VIII.

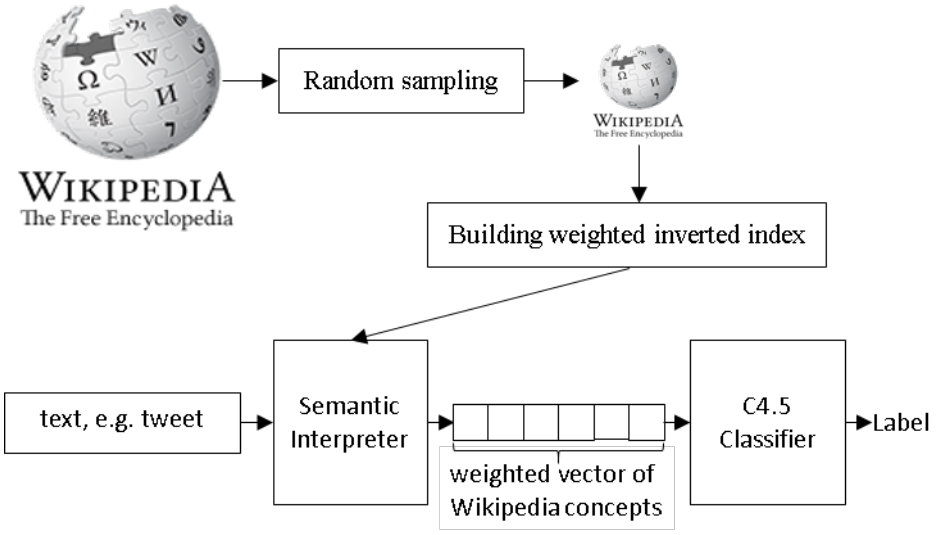

Figure 1. RS-ESA Overview

II. Randomized Explicit Semantic Analysis (RS-ESA)

As we mention in Section I, Explicit Semantic Analysis (ESA) is the state of the art approach for computing semantic relatedness, but its algorithm is very time-consuming due to the size of the Wikipedia dataset involved. At the moment of writing this publication, there are 4,857,074 articles in the English Wikipedia[1].

To improve the speed of the preprocessing step as well as the throughput of the ESA algorithm, we propose to utilize a sample of the Wikipedia dataset instead of the full dataset similar to the approach used to predict election results. In particular, it is impractical to ask everyone to make a decision and tally the ballots, which would produce 100% accurate results assuming honest answers. Instead, a sample of the population is interviewed in order to get results that reflect the target population as precisely as needed.

The level of precision in this case is affected by two parameters, namely confidence interval and confidence level. A confidence interval is a margin of error. For example, if a confidence interval is 2 and 95% of the sample picked a particular answer then we can be confident that the entire population would have picked that answer between 93% ($93-2$) and 97% ($95+2$). The confidence level indicates how sure we want to be. Depending on a problem various values can be utilized, but the most commonly used value is 95%. To determine the sample size for a proportion when sampling without replacement we can use the following equation from statistical inference:

$$n_0 = \frac{Z^2 p(1-p)}{\varepsilon},$$

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

where n_0 is the sample size without considering the finite population correction factor, Z or Z-score is a constant that represents the number of standard deviations a given proportion is away from the mean, p is the proportion and ε is the margin of error.

Applying the finite population correction factor results in the actual sample size n as follows:

$$n = \frac{n_0 N}{n_0 + (N-1)},$$

where N is the population size. Given Z-score=1.96 for 95% confidence level, N=4,857,074, and ε=0.02, the sample size n should be 2,400.

Our hypothesis is that a sample of the Wikipedia dataset can be used for the ESA method instead of the full dataset to improve its throughput while maintaining high precision. Recall that ESA represents the meaning of any text in terms of Wikipedia-based concepts. Concepts are the titles of Wikipedia articles characterized by the texts of those articles. In ESA a word is represented as a column vector in the TF-IDF table (table T) of Wikipedia concepts and a document is represented using its interpretation vector, which is a centroid of the column vectors representing its words. An entry $T[i,j]$ in the table of size $N \times M$ corresponds to the TF-IDF value of term t_i in document d_j, where M is the number of Wikipedia documents (articles) and N is the number of terms in those documents. See [18] for a more formal description of the ESA method.

For overview of RS-ESA approach - see Figure 1. Note, that we use a decision tree based classifier algorithm C4.5 in our experimental evaluation as we explain in Section VI.

III. Expert Based Explicit Semantic Analysis (Expert-ESA)

As we mention in Section II, we propose to use a random sample of Wikipedia concepts to speed up computations involved in explicit semantic analysis. As an alternative to this approach, we also investigate the use of a subset of Wikipedia repositories selected by an expert driven approach instead of random articles. This approach is thus tied to a particular domain being studied. In our case our domain is landslide detection and we are interested in classification of Social Media data as either relevant or irrelevant to landslide as a natural disaster.

The challenge here is that *landslide* is a polysemous word where one meaning is related to our domain and all other meanings are unrelated and represent noise as described in Section I. That is why we propose to extract a set of articles from Wikipedia that would represent meanings that are relevant to our domain, which is landslide as a natural disaster, and irrelevant meanings. In order to generate a set of articles that describe relevant and irrelevant meanings of our polysemous term, we propose the following approach. For both sets, we start with a list of initial Wikipedia concepts. Each of the Wikipedia articles representing those initial concepts contains links to other pages inside its text. The articles in these links are used as additional concepts for the corresponding sets. This process can be repeated multiple times to populate our sets of concepts. In this work we follow the links from each of the initial set of articles once and we demonstrate that the total number of concepts obtained this way is sufficient to label items with high precision in Section VI below.

To populate the set of relevant concepts, we can use the set of keywords used to collect landslide data from Social Media as our starting concepts. This set is represented by the following list of Wikipedia concepts: Landslide, Rockfall, Debris Flow, Mudflow, Flash Flood, Earthflow, and Rockslide. Each article representing these concepts contains a list of links to other articles that are also recorded. The total number of concepts extracted using this approach is equal to 550.

To populate the set of irrelevant concepts, we can use the set of Wikipedia concepts that represent the most common reasons for noise in Social Media with respect to landslide as a natural disaster, namely Landslide Victory, Blowout (sports), Landslide (song), Election, List of duo and trio cocktails. The last concept requires some explanation. There is no separate article in Wikipedia on the popular cocktail "Mudslide" as of writing this paper. However, there is an article listing several cocktails, including Mudslide, so we include that article into a list of irrelevant concepts. Similarly, each article representing these concepts contains a list of links that are also followed. The total number of irrelevant concepts extracted using this approach is equal to 716.

Social Media	Training Dataset	Evaluation Dataset
Twitter	26,953	42,268
YouTube	311	466
Instagram	136	204

Table I
Overview of dataset for landslide detection

Data Source	Evaluation Dataset	Class
Wikipedia articles	2,400	Factual
FanFiction Twilight Stories	2,400	Fictional

Table II
Overview of dataset of factual and fictional texts

IV. Implementation Details

A. Implementation Notes

To compute a sample size of the Wikipedia dataset for the RS-ESA approach we use the following values: population 4,857,074, confidence interval 2, confidence level 95%. The sample size based on the formulas listed in Section II is equal to 2,400. In order to select 2,400 random Wikipedia articles we first downloaded a list of all English page titles in main namespace from the Wikipedia dump dated March 4, 2015. Then we randomly selected a title from this list 2,400 times and downloaded a corresponding article using Wikipedia API[2].

Using this sampled dataset we generate table T where columns are titles of the Wikipedia articles, rows are all words present in those articles and $T[i, j]$ elements of the table are TF-IDF values. Note that we apply cosine normalization to each row to disregard differences in document length.

Next for each labeled text in the training and evaluation datasets we compute the centroid of the vectors representing the individual words. The centroid vectors of the training dataset are used to build classifier model, which is then used to predict labels for the centroid vectors of the evaluation dataset.

We perform classification analysis using the Weka software package [7]. Weka is an open source collection of machine learning algorithms and has become the standard tool in the machine learning community.

B. Processing Time

According to the authors of the original ESA approach, parsing of the Wikipedia XML dump on a standard workstation takes about 7 hours on a 2GHz dual core computer,

[2]https://pypi.python.org/pypi/wikipedia/

mostly due to the size of the entire Wikipedia corpus at that time. The number of articles in Wikipedia only increased since then. In our approach, the preprocessing step takes less than an hour on a comparable 2.67 GHz computer with 4 cores since we only use a sample of Wikipedia. Although it is a one-time operation, but its processing time still affects the applicability of the approach.

More importantly, the throughput of the original ESA approach is several hundred words per second, whereas RS-ESA's throughput is several thousand words per second, which is an order of magnitude improvement.

V. Description of Evaluation Datasets

We evaluate the performance of the proposed classification approach using two sets of data. The first dataset is based on the Social Media items collected for landslide detection purposes. The second dataset is based on the Wikipedia articles and FanFiction Twilight stories as sources for classifying texts into factual and fictional categories.

A. Datasets for Landslide Detection Using Social Media

The ground truth dataset for landslide detection includes both training and evaluation datasets - see Table I. The training dataset contains manually labeled items from Social Media, namely Twitter, Instagram and YouTube. In particular, it contains values from the *text* field for Twitter, values from the *caption* field for Instagram and values from the *title* and *description* fields for YouTube.

The data for the training dataset was collected during the period from August to December 2013. Labels are either *relevant* or *irrelevant* with respect to landslide as a natural disaster. To prepare a set of relevant items we need a list of confirmed landslides. For this purpose we use expert landslide publications. The USGS agency, in addition to earthquakes, also publishes a monthly list of landslide events collected from external reputable news sources, such as Washington Post, China Daily, Japan Times and Weather.com[3].

To find the Social Media items related to confirmed landslides within each month of the training period, we first filtered the data based on the landslide locations extracted from the confirmed landslides. Then we manually went through each item in the filtered list to make sure they described corresponding landslides by comparing the contents of the items with the corresponding landslide articles. And whenever there were URLs inside those social items, we looked at them also to make sure that they referred to the corresponding landslides. To create a list of unrelated items in the training set, we randomly picked items from each social source and manually went through each item. But this time we had to make sure that the items did not describe landslide events.

The data for the evaluation dataset was collected during the month of December 2014. Labels are again either *relevant* or *irrelevant* with respect to landslide as a natural disaster, but unlike the training dataset all geo-tagged items were labeled. Using the approach described for the training dataset, we identified all items related to the landslides reported by the USGS. Then we analyzed each of the remaining items and followed the URLs to confirm the candidate items' relevance to landslides. If the item did not contain a URL, then we tried to find confirmation of the described event on the Internet using its textual description as our search query. If another trustworthy source confirmed the landslide occurrence in the geo-tagged area then we marked the corresponding item as relevant. Otherwise we marked it as irrelevant.

For overview of data collection for landslide detection - see Figure 2. There is a separate downloading process based on the capabilities of each social network. But each downloading process uses the same set of landslide related keywords to retrieve data, including *landslide* and *mudslide*.

B. Dataset for Separation of Factual and Fictional Texts

The ground truth dataset for factual and fictional texts uses two input sources, namely Wikipedia articles and the FanFiction archive of Twilight stories[4]. We consider Wikipedia as a source of factual data and Twilight stories as a source of fictional data.

Our ground truth dataset contains 2,400 Wikipedia articles and 2,400 fan fiction stories. To randomly select 2,400 Wikipedia articles, we again used a list of all English page titles in main space. Then we randomly selected a title from this list 2,400 times. We applied a similar approach to randomly select 2,400 fan fiction stories. First we downloaded 41,851 stories from the FanFiction archive. Note, that we only downloaded the first page of each story to speed up the downloading process. Then we randomly selected an article from this list 2,400 times making sure that the article contained at least 100 words.

For overview of data collection for separation of factual and fictional texts - see Figure 3. The labeling process here does not require user input, because we automatically label all Wikipedia articles as *factual* and all FanFiction stories as *fictional*. The experimental evaluation of separation of factual and fictional texts uses 10-fold cross-validation approach, so there is a single evaluation dataset.

VI. Experimental Evaluation

In this section we present an experimental study of the proposed RS-ESA approach and compare it with the Expert-ESA approach. We designed 4 sets of experiments for evaluation purposes. We start by analyzing the effectiveness of RS-ESA for identifying relevance of Social Media data

[3]http://landslides.usgs.gov/recent/

[4]https://www.fanfiction.net/book/Twilight/?&srt=1&r=103&p=1

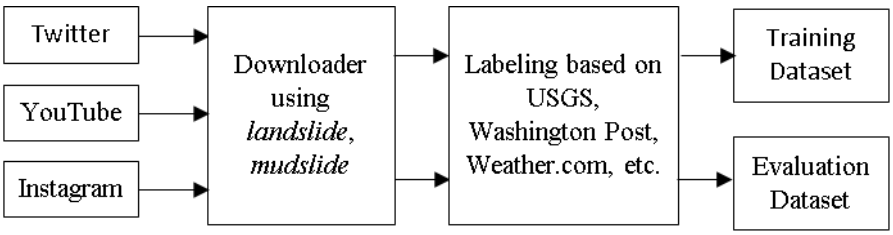

Figure 2. Overview of Data Collection for Landslide Detection

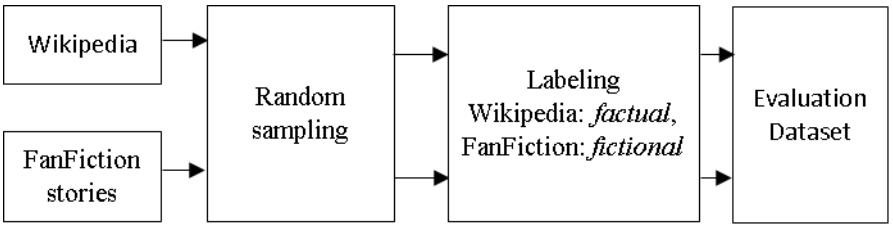

Figure 3. Overview of Data Collection for Factual and Fictional Texts

to landslide as a natural disaster using a random sample of Wikipedia repository. To confirm our results we generate a second random sample of Wikipedia repository and perform evaluation of landslide classification again. Next we evaluate Expert-ESA approach and run a third classification analysis of landslide data. Finally, we use RS-ESA approach to perform classification analysis of separating factual texts from fictional.

Note, that we do not include comparison of classification results based on RS-ESA and Expert-ESA approaches versus original ESA, because we were unable to compute a semantic interpreter using the latest Wikipedia XML dump within a reasonable amount of time. However, we intend to add comparisons of both RS-ESA and Expert-ESA versus baseline methods, such as Bag-of-Words approach and others, as part of our future work.

A. Classification of Social Media for Landslide Events

In this and all other experiments we use a decision tree based classifier algorithm C4.5. We choose it, because we want a classifier algorithm to reflect the process of how we built the ground truth dataset for landslide detection described in Section V. In particular, during the process of manually labeling items from Social Media we noticed that we could almost instantly tell whether a given social item was relevant to landslide as a natural disaster or not. There are several common both relevant and irrelevant topics discussed in Social Media that are easy to spot due to the use of specific words. Each time a particular word was used we could predict with high accuracy the label of the whole text. Hence, we choose a decision tree based algorithm that predicts labels based on the thresholds of the relevance of terms to the concepts represented as features. Note, that Weka's implementation of the C4.5 algorithm is called J48.

For the first experiment we first generated a random sample of 2,400 Wikipedia articles, including:

- Title 1: "Marquetry"
- Title 1,200: "Chemokine receptors"
- Title 2,400: "Shah Kalim Allah Jahanabadi"

Next we generated table T using the words from these articles as rows, titles as columns and the corresponding normalized TF-IDF values as elements. Using this table we computed the centroid vectors for both training and evaluation datasets. Next we used Weka to build a classifier model based on the centroid vectors from the training dataset. Using this model we classified the centroid vectors from the evaluation dataset. For results of classification performance using this approach see row RS-ESA 1 in Table III. Note, that in spite of a rather low recall of 66% precision is very high at 97%.

To validate high precision results we generated another random sample of 2,400 Wikipedia articles, including:

- Title 1: "980 African Cup of Nations Final"
- Title 1,200: "Macrocneme nigritarsia"
- Title 2,400: "Paleontology in Utah"

For results of classification performance using this sample see row RS-ESA 2 in Table III. Note, that although precision is a little lower, but it is still quite high at 96%, while recall is higher at 78% and F-score exceeded 86%.

Next we evaluate classification performance using Expert-RSA approach described in Section III. Using the related Wikipedia articles downloaded according to the described method, we generated a new table T using the same approach. Using this table we computed the centroid vectors for both training and evaluation datasets for landslide detection. For results of classification performance using this approach see row Expert-RSA in Table III. As expected,

explicit semantic analysis based on a set of articles selected using an expert driven approach, had a better performance. However, this method requires manual initialization of the starting concepts used to download related articles by an expert user. Also, classification precision achieved using RS-ESA approach is quite high, while not requiring an input from user. It should also be noted that recall using RS-ESA approach is inferior to Expert-ESA, which is why we plan to continue improving RS-ESA performance.

B. Classification of Factual and Fictional Texts

Our final experiment is designed to evaluate the genericity of the RS-ESA approach by classifying data from a different domain. In particular, we choose the problem of classifying texts into factual and fictional categories. For this purpose we use a popular archive of fan fiction, in particular Twilight stories [5]. We consider Wikipedia as a source of factual data and Twilight stories as a source of fictional data.

We reuse the table T generated for RS-ESA 1 experiment. Our evaluation dataset for this experiment is described in Section V. We compute the centroid vectors for texts in the evaluation dataset using table T and assign label *factual* to Wikipedia articles and label *fictional* to Twilight stories. We apply 10-fold cross validation using C4.5 classifier and obtain a high value of precision again at 97%.

Approach	Precision	Recall	F-score
RS-ESA 1	97%	66%	79%
RS-ESA 2	96%	78%	86%
Expert-ESA	98%	84%	91%

Table III
CLASSIFICATION OF LANDSLIDE EVENTS

VII. RELATED WORK

Text classification (also known as text categorization, or topic spotting) is used to automatically sort a set of documents into classes (or categories, or topics) from a predefined set [8]. It has attracted a booming interest from researchers in information retrieval and machine learning areas in decades. Recently, several novel classification approaches have been proposed and implemented in text classification. Pu Wang et al. [9] presented semantics-based algorithm for cross-domain text classification using Wikipedia based on co-clustering classification algorithm. Elisabeth Lex et al. [10] described a novel and efficient centroid-based algorithm Class-Feature-Centroid Classifier(CFC) for cross-domain classification of web-logs, also they have discussed the trade-off between complexity and accuracy. Pan et al. [11] proposed a spectral feature alignment (SFA) algorithm to align domain-specific words from different domains into unified clusters, with the help of domain independent words as a bridge. Zhen et al. [12] propose a two-stage algorithm which is based on semi-supervised classification to address the different distribution problem in text classification.

ESA was first introduced by Gabrilovich et al. [6] as an approach to compute the semantic relatedness of terms or short phrases. Since then, lots of researchers have used ESA in many applications successfully. Egozi et al. [13] used ESA for the estimation of the relevance of documents for a given query and selected high quality features for classification. Potthast et al. [14] and Sorg et al. [15] proposed a cross-lingual extension (CL-ESA) that exploits interlanguage links of Wikipedia articles. Cimiano et al. [16] presented that CL-ESA is superior to other retrieval models which are based on implicit semantics. Also, ESA is used to compute the semantic relatedness of terms. For instance, Mller et al. [17] used ESA as parameters in other retrieval models. In addition, several studies have been conducted to understand or enhance ESA performance ([18]). Anderka and Stein revisited ESA and found syntactic parallels to the generalized vector space model (GVSM [19]). They also conducted some initial analysis targeting the impact of the index collection on the performance of ESA. They concluded that the ESA is a general methodology that can be applied on any corpus with concept-level titles or categories. We focus on the Wikipedia use here following several other studies [20], [21]. These studies mostly use Wikipedia corpus to generate concept vectors, and therefore the resulted vector is a vector of Wikipedia concepts given a text document. For example, Scholl et al. (2010) proposed enhancements to ESA (called Extended Explicit Semantic Analysis) that make use of further semantic properties of Wikipedia like article link structure and categorization, thus utilizing the additional semantic information that is included in Wikipedia.

Text mining has been widely used in detection systems for disaster events such as earthquakes and hurricanes. Sakaki et al. [1] proposed an algorithm to monitor tweets and detect earthquake events by considering each Twitter user as a sensor. Cameron et al. [22] developed platform and client tools called Emergency Situation Awareness - Automated Web Text Mining (ESA-AWTM) system by identifying tweets relevant to emergency incidents. Wang et al. [23] proposed a mixture Gaussian model for bursty word extraction in Twitter and then employed a novel time-dependent HDP model for new topic detection. Hua et al. [24] presented STED, a semi-supervised system that helps users to automatically detect and interactively visualize events of a targeted type from Twitter, such as crimes, civil unrests, and disease outbreaks. Our previous work LITMUS [3], [4], [5] adopts text mining techniques for data analysis on data from multiple information sources such as physical and social information services. To achieve optimized performance of

[5]https://www.fanfiction.net/book/Twilight/?&srt=1&r=103&p=1

the detection system in terms of precision, we have spent lots of research efforts on improving the text mining techniques in general.

VIII. Conclusion

Automated text classification or text categorization is an important problem in computer science. In this paper we propose a new approach for fast text classification based on randomized explicit semantic analysis (RS-ESA), whose throughput is an order of magnitude faster than the original explicit semantic analysis approach. We demonstrate that our approach using a random sample of Wikipedia articles achieves 96% precision when classifying Social Media landslide data collected in December 2014. We compare the results achieved using RS-ESA approach with explicit semantic analysis approach based on a subset of Wikipedia articles selected by an expert (Expert-ESA) next. Finally, we demonstrate the genericity of the proposed RS-ESA approach by successfully applying it to a different problem where we separate factual texts from fictional based on Wikipedia articles and fan fiction stories, where we achieve 97% precision.

Due to promising results achieved in separating factual texts from fictional using RS-ESA approach based on a limited number of texts, we intend to expand our tests by increasing the size of the evaluation dataset as part of the future work. We plan to add more kinds of sources of factual and fictional texts to confirm our results in diverse domains. We are also interested in evaluating the influence of the sample size on classification performance. Similarly, we are interested in evaluating the influence of the selected concepts used to build ESA table. We plan to run our method multiple times and report average performance achieved. Finally, we intend to evaluate both Expert-ESA and RS-ESA approaches in other domains.

Acknowledgements

This research has been partially funded by National Science Foundation by CNS/SAVI (1250260, 1402266), IUCRC/FRP (1127904), CISE/CNS (1138666, 1421561) programs, and gifts, grants, or contracts from Fujitsu, HP, Intel, Singapore Government, and Georgia Tech Foundation through the John P. Imlay, Jr. Chair endowment. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation or other funding agencies and companies mentioned above.

References

[1] T. Sakaki *et al.*, "Earthquake shakes twitter users: real-time event detection by social sensors," in *WWW*, 2010.

[2] E. Miltsakaki *et al.*, "Real-time web text classification and analysis of reading difficulty," in *Proceedings of the Third Workshop on Innovative Use of NLP for Building Educational Applications*, ser. EANL '08, 2008, pp. 89–97.

[3] A. Musaev *et al.*, "LITMUS: Landslide Detection by Integrating Multiple Sources," in *ISCRAM*, 2014.

[4] A. Musaev *et al.*, "Landslide detection service based on composition of physical and social information services," in *Web Services (ICWS), 2014 IEEE International Conference on*, June 2014, pp. 97–104.

[5] A. Musaev *et al.*, "LITMUS: a Multi-Service Composition System for Landslide Detection," *IEEE Transactions on Services Computing*, vol. PP, no. 99, 2014.

[6] E. Gabrilovich *et al.*, "Computing semantic relatedness using wikipedia-based explicit semantic analysis," in *Proceedings of the 20th International Joint Conference on Artificial Intelligence*, ser. IJCAI'07, 2007, pp. 1606–1611.

[7] M. Hall *et al.*, "The weka data mining software," *ACM SIGKDD Explorations Newsletter*, vol. 11, no. 1, 2009.

[8] F. Sebastiani, "Text categorization," in *Text Mining and its Applications to Intelligence, CRM and Knowledge Management*, 2005.

[9] P. Wang *et al.*, "Cross-domain text classification using wikipedia," *IEEE Intelligent Informatics Bulletin*, vol. 9, no. 1, 2008.

[10] E. Lex *et al.*, "Efficient cross-domain classification of weblogs," *International Journal of Intelligent Computing Research*, vol. 1, no. 1, 2010.

[11] S. J. Pan *et al.*, "Cross-domain sentiment classification via spectral feature alignment," in *WWW*, 2010.

[12] Y. Zhen *et al.*, "Cross-domain knowledge transfer using semi-supervised classification," in *AI 2008: Advances in Artificial Intelligence*, 2008, vol. 5360.

[13] O. Egozi *et al.*, "Concept-based feature generation and selection for information retrieval," in *Proceedings of the 23rd National Conference on Artificial Intelligence - Volume 2*, ser. AAAI'08, 2008, pp. 1132–1137.

[14] M. Potthast *et al.*, "A wikipedia-based multilingual retrieval model," in *Proceedings of the IR Research, 30th European Conference on Advances in Information Retrieval*, ser. ECIR'08, 2008, pp. 522–530.

[15] P. Sorg *et al.*, "Cross-lingual information retrieval with explicit semantic analysis," in *Working Notes for the CLEF 2008 Workshop*, ser. CLEF 08, 2008.

[16] P. Cimiano *et al.*, "Explicit versus latent concept models for cross-language information retrieval," in *Proceedings of the 21st International Jont Conference on Artifical Intelligence*, ser. IJCAI'09, 2009, pp. 1513–1518.

[17] C. Mller *et al.*, "Semantically enhanced term frequency," in *Advances in Information Retrieval*, ser. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2010, vol. 5993, pp. 598–601.

[18] M. Anderka *et al.*, "The esa retrieval model revisited," in *Proceedings of the 32Nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '09, 2009, pp. 670–671.

[19] S. K. M. Wong *et al.*, "Generalized vector spaces model in information retrieval," in *Proceedings of the 8th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '85, 1985, pp. 18–25.

[20] Z. Minier *et al.*, "Wikipedia-based kernels for text categorization," in *Symbolic and Numeric Algorithms for Scientific Computing, 2007. SYNASC. International Symposium on*, Sept 2007, pp. 157–164.

[21] P. Scholl *et al.*, "Extended explicit semantic analysis for calculating semantic relatedness of web resources," in *Sustaining TEL: From Innovation to Learning and Practice*, ser. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2010, vol. 6383, pp. 324–339.

[22] M. A. Cameron *et al.*, "Emergency situation awareness from twitter for crisis management," in *WWW Companion*, 2012.

[23] X. Wang *et al.*, "Real time event detection in twitter," in *Web-Age Information Management*, 2013, vol. 7923.

[24] T. Hua *et al.*, "STED: semi-supervised targeted-interest event detectionin in twitter," in *KDD*, 2013.