Systems Engineering Methods Business and Analytics

Technical Report SERC-2019-TR-016

November 15, 2019

Principal Investigator: Dr. K.P. Subbalakshmi, Stevens Institute of Technology

Co-Principal Investigator: Dr. William Rouse, Stevens Institute of Technology

Research Team:

Stevens Institute of Technology: Harish Sista

Sponsor: US Army RDECOM-ARDEC; ODASD(SE)

The Systems Engineering Research Center (SERC) is a federally funded University Affiliated Research Center managed by Stevens Institute of Technology.

This material is based upon work supported, in whole or in part, by the U.S. Department of Defense through the Office of the Assistant Secretary of Defense for Research and Engineering (ASD(R&E)) under Contract [HQ0034-19-D-0004, TO#0495].

1 TABLE OF CONTENTS

2 LIST OF FIGURES

3 LIST OF (TABLES, SEQUENCES)

4 GOAL AND EXECUTIVE SUMMARY

The goal of this project is to develop and compare natural language processing (NLP) based tools that will extract keyphrases from scientific literature of interest to CCDC. Keyphrase extraction is an important first step in several downstream NLP tasks including summarization, opinion mining, trend analysis etc.

We start with the CCDC SME provided keyphrases listed in Table-1.

Table 1: Keyphrases provided by CCDC

Additive technology
Chemical additive to solutions
Cryogenic milling
Microfluidics
Nanopowder
Nanoscale

The project was executed in the following steps:

- Extract meaningful datasets from scientific literature relevant to CCDC
- Extract topics using Latent Dirichlet Allocation (LDA) methods from these datasets
- Use this as a base and add on a couple of keyphrase extraction methods to extract keyphrases. Specifically we used Rapid Automatic Keyword Extraction algorithm(RAKE) [1] and Position Rank Analysis (PRA) [4].
- Compare the LDA-RAKE, LDA-PRA and PRA on the datasets obtained.
- Stress test these algorithms using standard datasets like the NUS dataset.

The technical details of all the methods and component methods are available in Section 2.1.1 . The methods of scoring algorithms are described in Section 2.1.2. The description of all datasets is available in 2.1.4. Finally the results are described in Section 2.1.3 and conclusions appear in Section 2.1.5.

4.1.1 METHODS AND ALGORITHMS:

We start the project with determining topics in the paper. For this we use the Latent Dirichlet Allocation (LDA) [1]. LDA expresses these topics as a probability distribution of keywords. Since the goal is to extract keyphrases, rather than probability distributions, we built two other models on top of LDA and tested it with another existing keyphrase extraction method called position rank analysis (PRA) [4]. These algorithms are described in detail in this section.

4.1.1.1 Latent Dirichlet Algorithm (LDA)

LDA is a statistical topic-based model [1]. This model is used for generating topics from a given text corpus. LDA assumes that every text corpus(here in our case research paper) is a probability distribution of the topics associated with it, and every topic itself is a probability distribution of the keywords associated with it.
Given the text data(research paper) and the number of topics that are to be generated from it the LDA model generates the specified number of topics and keywords that are associated with these topics. Each sentence in the paper can be represented as probability distribution of these topics, that is why the paper is broken down to individual sentences and the LDA analysis is performed on the sentences of the paper.
Given a paper D with the collection of sentences W and the number of topics K. LDA assumes that the paper is a Dirichlet distribution of topics(Z_K) where each topic is a probability distribution of keywords of the paper. The Dirichlet distribution of the paper is defined on a (K-1) simplex which is a topics plane(Z_K) with a Dirichlet random variable θ. The probability density function for the random variable θ is given as follows, where Γ(x) is the Gamma function.

$$p(\theta \mid \alpha) = \frac{\Gamma\left(\sum_{i=1}^{k} \alpha_i\right)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \theta_1^{\alpha_1 - 1} \cdots \theta_k^{\alpha_k - 1}$$

The probability of every keyword (W_j) for every topic(Z_K) can be represented in a matrix β where $\beta_{ij} = p(W_j = 1 \mid Z_i = 1)$. Given these parameters α and β, the research paper can be represented as a joint probability distribution which is given as follows.

$$p(\theta, \mathbf{z}, \mathbf{w} \mid \alpha, \beta) = p(\theta \mid \alpha) \prod_{n=1}^{N} p(z_n \mid \theta) p(w_n \mid z_n, \beta)$$

Since the dataset already contains the research paper, the parameters α and β are estimated using maximum-likely hood estimation, and buy using the estimated parameters α and β the Topics and the topic keywords for the papers are estimated.
To generate this probabilistic model LDA only requires to check for the occurrence of the keyword in a sentence. So, it doesn't matter where the keywords are positioned in the sentence, hence the Bag of Words representation of the sentences has been used. Also, since there are other words in a text document which are more frequently repeated and doesn't hold any semantic meaning, they are categorized as Stop-Words and are filtered from the sentences. Thus, the obtained Bag of Words from each sentence are given as input for LDA model. The Gensim's LDA model and the Stop-words library [2] have been used for generating the Bag of words representation of the sentences and to implement the LDA topic model.

4.1.1.2 Rapid Automatic Keyword Extraction (RAKE) Algorithm

RAKE [3] is a frequency-based model used for extracting keywords and key-phrases from short passages. Using the co-word frequency between the key-words this algorithm generates the keyphrases which can explain more about the semantic relation between the keyphrases (e.g., "Joan of Arc", "Joan and Arc"). As a keyphrase generating model this method does filer stop words from the text, but the model does not filter the stop words which occur between the keywords with high co-word frequency. Thus, the generated keyphrases are taken as collective entities along with other keywords inside a sentence. To generate the keyphrases first the co-word matrix of all the keywords in the sentence is generated and the following metrics for all each keyword, w, are calculated.

- freq(W) → The number of times the word has occurred in the sentence. Called the frequency
- deg(W) → The number of times the keyword has occurred adjacent to other keywords. Called the degree.
- (deg(W)/freq(W)) → ratio of degree to frequency

This algorithm assumes that every word other than the stop words is a keyword to begin with It then calculates the above metric for each keywords. Keywords that have high ratio of degree to frequency are combined together to form the keyphrases. If there are any stop words between these keywords, they are not filtered and included between the keywords to generate the keyphrases with stop words. This is one of the earliest keyphrase generating mechanisms.

4.1.1.3 LDA-RAKE

The first model we constructed to extract keyphrases from a document was to combine LDA with RAKE. RAKE extracts keywords and keyphrases, but only from one sentence at a time. If this were to be directly used in a document, then it might result in a set of keyphrases for each sentence in the document. For a scientific document, this could mean a lot of keyphrases.

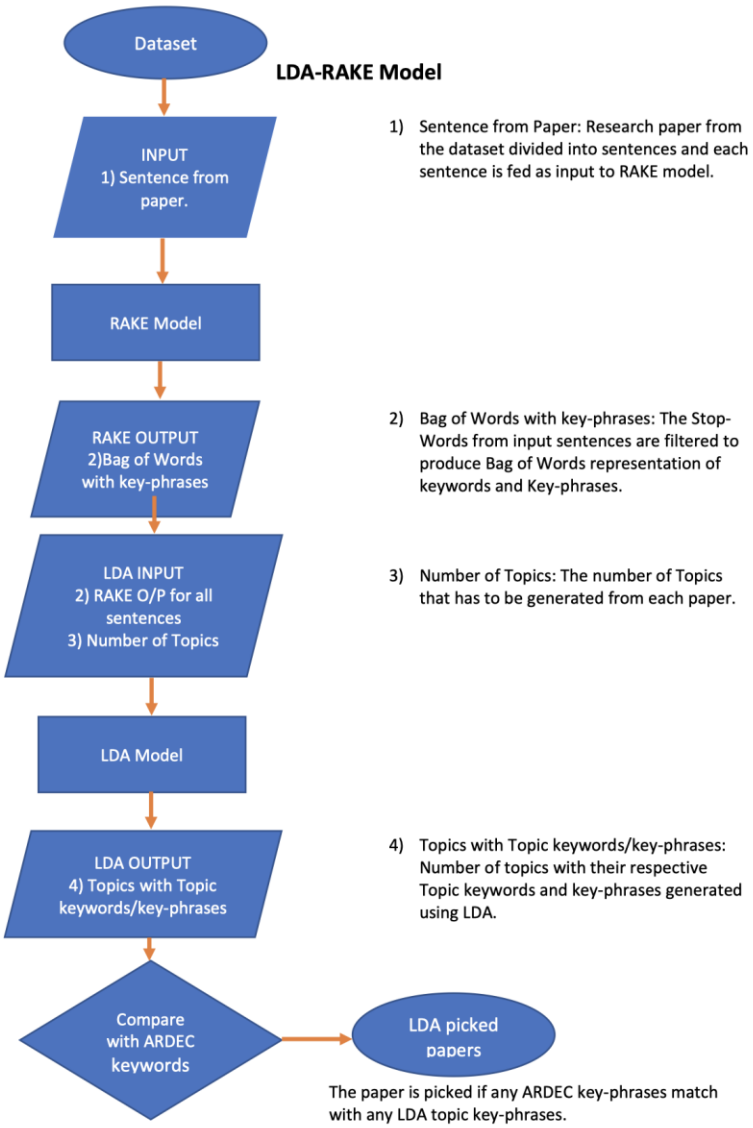

Figure 1: LDA-RAKE Architecture

As mentioned earlier, LDA generates topic distributions for a document. It uses a bag-of-words as its input. For LDA-RAKE, we first applied RAKE to each of the sentences. This was then fed to the LDA to create topic distributions, which now consists of keywords and keyphrases. The flowchart for LDA-RAKE model is as shown in the Figure 1.

The LDA-RAKE model was executed on every paper from the CCDC dataset. As the first step every sentence from each paper is processed with RAKE to generate keywords and keyphrases. In the second step the output of RAKE for each paper is processed with LDA model to generate topics with keywords and keyphrases. In the final step these LDA-RAKE extracted topic keywords and keyphrases for each paper are compared with the subject matter expert (SME) keyphrases, if there is at least one match found between the topic keyphrases and the SME

keyphrases then the paper is tagged as one of interest to the SME. Note that LDA-RAKE generate repeated keyphrases. These are manually deconflicted, so that there is no repetition in the keyphrases.

The results of the LDA-RAKE for CCDC dataset are listed in Table 2. Ten topics were extracted from each paper. LDA-RAKE has identified a total of 6 papers, the names of the papers are mentioned in the "identified papers" section and all the topic keywords and keyphrases extracted from the 10 topics are mentioned in the topic keyphrases section. The analysis of the topic keywords and keyphrases shows that, the number of key phrases that are generated are relatively high to find he trending topics from the selected papers, there are some keywords like 'results', 'contain' and 'varying' which doesn't hold any semantic meaning that can describe the document also there are keywords like 'tion' which are generated because of the pdf to text conversion tool. The conclusion is even though there are many key phrases generated for each paper, the quality of the keyphrases is not good enough to be selected for trending topics.

Table 2: Auto-tagged papers by LDA-RAKE and the keyphrases

LDA-RAKE identified papers	LDA-RAKE topic keyphrases
Combustion of Nanoaluminum and Water Propellants: Effect of Equivalence Ratio and Safety/Aging Characterization	'likely due', 'work', 'burning rate', 'figure', 'presence', 'witness plate', 'risha r', 'varying', 'tested', 'results', 'propellant sample', 'likely', 'full paper', 'r sippel', 'aluminum nanopowders', 'flame temperature', 'nanoparticle agglomeration', 'pfeil', 'ciency', 'nanoaluminum', 'risha', 'combustion', 'placed', 'exponent', 'chamber', 'water propellants', 'petn', 'nova', 'propellants', 'nanoscale aluminum', 'l alex', 'role', 'lunar', 'fuel', 'performance', 'wt al wt al', 'composition', 'frozen water', 'f son', 'insensitive', 'risha et al', 'reaction', 'micro', 'liquid', 'al h', 'freezing', 'argonide', 'stoichiometric', 'possible', 'water', 'pressure', 'alumina', 'shock', 'inst', 'thermal conductivity', 'aiaa', 'alex', 'diffusion limitation', 'shown', 'psia', 'combustion efficiency', 'school', 'glomeration', 'embedded', 'aluminum', 'cliff et al', 'aluminum combustion', 'propellants explos pyrotech', 'transition', 'dependence', 'nal variety', 'effect', 'equivalence ratio', 'ignited', 'frozen nal h', 'tech', 'ignition', 'contain', 'www pep wiley vch de', 'yetter', 'tion', 'l pourpoint', 'combus', 'propellant', 'initial'
Multi-Parameter Study of Nanoscale TiO 2 and CeO 2 Additives in Composite AP/HTPB Solid Propellants	'nano dry', 'composite ap htpb solid propellants', 'burning rate', 'wiley vch verlag gmbh co kgaa weinheim', 'additive type', 'baseline', 'sensitivity', 'propellants explos pyrotech', 'produced', 'performance', 'full paper', 'analyzing', 'powder', 'seal', 'stephens e l petersen r carro', 'l reid', 'size', 'provided', 'nano wet', 'results', 'additive', 'stephens', 'nano', 'reported', 'fact', 'tailoring', 'sequence', 'nanoscale tio', 'nano sized', 'nanoparticles', 'additive percentage', 'test', 'formula', 'effectiveness', 'mixing method', 'method', 'achieve', 'ingredient', 'development', 'ceo additives', 'would like', 'micro dry', 'solvent', 'current study', 'agglomeration', 'pressure exponent', 'final propellant grain', 'effect', 'ammonium', 'water', 'pressure', 'properly', 'hand mixing method', 'effects', 'vacuum', 'mixing', 'burning rates', 'oxidizer', 'ceria', 'rutile', 'propulsion conference', 'usa july aiaa paper', 'aiaa asme sae asee joint propulsion conference', 'tested', 'multi parameter study', 'figure tem image', 'study', 'analysis', 'summary', 'inhibiting', 'figures', 'evident', 'discussion', 'www pep wiley vch de', 'tion', 'propellant', 'additives'
Trinitrotoluene Nanostructuring by Spray Flash Evaporation Process	'diethyl ether', 'sprayed', 'wiley vch verlag gmbh co kgaa weinheim', 'nozzle', 'lization', 'h leubner', 'growth', 'acetaldehyde', 'concentration', 'technology', 'size', 'trinitrotoluene nanostructuring', 'tnt structure', 'v pichot f schnell', 'particles', 'rapid expansion', 'role', 'acetone', 'samples', 'utions', 'vaporization', 'synthesis', 'product', 'mixture', 'solvent', 'spitzer', 'structure', 'influence', 'practice', 'controlling crystal size eds crc press pp', 'mtbe', 'nanoscale', 'sfe process', 'spray flash evaporation process', 'present', 'case', 'propellants explos pyrotech', 'full paper', 'solution', 'well known', 'crystal', 'ticles', 'nanocrystalline rdx', 'tnt hns', 'www pep wiley vch de', 'precision crystallization theory'

Transition from Impact-induced Thermal Runaway to Prompt Mechanochemical Explosion in Nanoscaled Ni/Al Reactive Systems	'hebm material', 'powder mixtures j appl phys', 'concept', 'appl phys lett', 'gordopolov', 'wiley vch verlag gmbh co kgaa weinheim', 'appeared', 'slow reaction', 'v f nesterenko', 'f x jette', 'radulescu j j lee', 'thermal', 'nanopowder', 'stress wave passage', 'impact event', 'speed', 'normal stress', 'impact', 'slow', 'temperature', 'passage', 'varma', 'reactive', 'charron tou', 'shear', 'initiated', 'entire', 'material', 'possible', 'l gur ev', 'bulk', 'heat', 'crush strength', 'high speed reaction mode', 'plunger', 'sample', 'structural changes', 'goroshin', 'impact event ini', 'j higgins', 'ignition', 'www pep wiley vch de', 'tion', 'mte mode', 'hibited', 'figure', 'shock compression', 'aip conference proceedings', 'reduce', 'tion mode', 'high', 'rate', 'full paper', 'compression', 'plastic deformation', 'shown', 'angle', 'reaction behavior', 'able', 'crush', 'sound speed', 'levitas v f nesterenko', 'signant', 'contrast', 'mode', 'also shown', 'investigated', 'ples', 'meyers strain induced', 'american physical society topical group', 'small', 'produce', 'batsanov', 'reaction', 'notre dame', 'l frost', 'chemical reaction', 'prompt reaction', 'materi', 'heating', 'materials', 'situ measurements', 'time', 'approximate location', 'phys chem c'
Studies on the Non-isothermal Crystallization Behavior of Aluminum Nanopowder-Filled Poly(3,3-bis-azidomethyl Oxetane)	'slope', 'figure', 'c min', 'ict karlsruhe germany june july p', 'calculated', 'rate', 'full paper', 'yunjun luo kai guo', 'plot', 'shown', 'temperature', 'introduction', 'crystallization', 'dhc pbamo', 'approach', 'synthesis', 'alumi', 'constant cooling rate', 'composite', 'respectively', 'formed', 'effect', 'ozawa exponent', 'cooling rate', 'non isothermal crystallization behavior', 'data', 'non isothermal crystallization', 'aluminum nanopowder filled poly bis azidomethyl oxetane', 'wiley vch verlag gmbh co kgaa weinheim', 'straight line', 'aluminum nanopowder', 'ln k', 'ict karlsruhe', 'www pep wiley vch de', 'ln b', 'propellants explos pyrotech'
Formulation and Characterizations of Nanoenergetic Compositions with Improved Safety	'phenomenon', 'wiley vch verlag gmbh co kgaa weinheim', 'nanocompositions', 'wuillaume', 'observed', 'tions', 'impact h mm pmax mpa', 'sample', 'improved safety', 'wt rdx nanocryogel', 'full paper', 'performed', 'formulation', 'g per batch scale', 'figure sem image', 'matrix', 'nano', 'density', 'propulsion', 'key point', 'classical', 'charge', 'experiment', 'decomposition', 'critical diameter', 'expected', 'figure', 'ssgt', 'composition', 'nanoscale', 'nanoenergetic compositions', 'macro', 'solvent', 'beaucamp f david quillot c erad', 'reduce', 'impact sensitivity', 'crystal quality', 'possible', 'p np f rdx', 'p np f ap', 'macrocomposi', 'freeze drying', 'pellet', 'safety', 'improve', 'case', 'stability', 'carried', 'pressure', 'rdx wt', 'propellants explos pyrotech', 'resin', 'g cm', 'matter', 'residual po', 'pressing sequence', 'www pep wiley vch de', 'g butyrolactone', 'nano rdx', 'order'

4.1.1.4 Position Rank analysis(PRA)

Position Rank Analysis [4] is an unsupervised frequency-based model used for calculating the keyphrases based on ranks from a text corpus. Unlike RAKE, PRA calculates the keyphrases based how frequently they keywords are occurring in consecutive sentences, the co-occurrence score(PageRank Score) of the keywords is scored based on the position of the sentence in the document.

Say if two keywords (k1,k2) co-occur in the beginning of the sentence of the paper and the keywords (t1,t2) co-occur at the ending sentence of the paper. Then the PRA model gives higher PageRank score to (k1,k2) and less PageRank score to (t1,t2). Also, the PRA model is designed to find the keywords which follow only a given pattern of parts of speech structure. Hence the parts of speech tags of sentences are used in finding the keyphrases.
The infrastructure of original PositionRank analysis [4] can be explained in three parts.

- Graph Construction at word level
- Design of Position biased page rank.
- Formation of Candidate Phrases.

Graph Construction at word level:

Given a document "D", a graph G=(V, E) is constructed in such a way that the V is the set of all the Nouns and Adjectives that are extracted from document D. Then, an E_{ij} is is defined as the edge connecting two nodes $\{V_i, V_j\}$. The weight of an edge E_{ij} between two nodes V_i and V_j is defined as the number of times these two nodes have co-occurred in the document "D" with a window of size "w".

Design of Position biased page rank:

The PageRank score [5] is the importance rank of each vertex V_i within the graph G. The PageRank for each vertex V_i is calculated by recursively summing up the normalized weights of edges at every step. Let "M" be the adjacency matrix of all nodes V, then the initial page rank for a node V_i is set to 1/|V| where |V| is the initial norm of all vertices and then it is incremented according to the formula shown below.

$$S(t+1) = \widetilde{M} \cdot S(t)$$

Where S(t) is the Score of all vertices after processing window of size "w" at state "t", $M^{\sim}$ is the normalized form of the matrix M, each element $m_{ij}^{\sim}$ of the matrix $M^{\sim}$ is defined as follows.

$$\widetilde{m_{ij}} = \begin{cases} m_{ij} / \sum_{j=1}^{|V|} m_{ij} & \text{if } \sum_{j=1}^{|V|} m_{ij} \neq 0 \\ 0 & \text{otherwise} \end{cases}$$

Thus, these nodes and steps can be represented as states and transitions of Markov's chain. By recursively performing the above procedure, the principal eigenvector representing

all states is calculated. The principal eigenvector matrix S is given as follows where α is the damping factor(to not to get stuck in loops in states) and p is a vector of normalized probabilities of all elements.

$$S = \alpha \cdot \widetilde{M} \cdot S + (1 - \alpha) \cdot \widetilde{p}$$

The PositionRank model considers that the words that are mentioned at the beginning of a scholarly document ted to repeat more frequently throughout the document, that is why the words that appear at the beginning of the document are given higher PageRank scores as more steps are introduced.

Candidate Key-Phrase Generation:

The keywords which are adjacent and show contiguous key-phrase scores and satisfy the equation (adjective)*(noun)+ are concatenated as key-phrases. The weights for the individual word-vectors are added to form the weights for the key phrases. The length for these key-phrases can go up to a length of three.

The flow chart of PRA analysis is as shown in the Figure 2. The PRA analysis is performed on every paper from CCDC dataset. Since PRA is strictly based on position of the keywords in the document, in the first step the text is split based on window size and the PRA analysis is performed on these windows. In the second step unlike LDA instead of filtering the stop words, in PRA the Nouns and Verbs from the windows are extracted using Parts of Speech(POS) tags. The output of the Parts of Speech(POS) analysis, is fed to PRA analysis to extract the ranked keywords for every paper. Only the top 10 ranked keyphrases are extracted for each paper in the CCDC dataset, if at least one match is found between the extracted keyphrases and the SME keyphrases then the paper is chosen to be related to the problem of interest.

Table 3: Auto-tagged papers using PRA and the corresponding keyphrases

PRA Identified papers	PRA Keyphrases
Fabrication and Properties of Insensitive CNT/HMX Energetic Nanocomposites as Ignition Ingredients	Cnt_hmx_nanocomposites, cyclotetramethylene_tetranitramine_hmx, cnt_hmx, composite_hmx_particle, tetranitramine_hmx, cnt_cnt, hmx_ure_show, hmx_particle_increase, nanoscale_hmx_particle, hmx_particle
Recent Developments for Prediction of Power of Aromatic and Non-Aromatic Energetic Materials along with a Novel Computer Code for Prediction of Their Power	Explosive_power_measurement, explosive_power, energetic_compound_show, several_energetic_compound, desired_energetic_compound, specific_energetic_compound, many_energetic_compound, pure_energetic_compound, energetic_compound_includ, nanoscale_energetic_compound

Nano Aluminum Energetics: The Effect of Synthesis Method on Morphology and Combustion Performance	Particle_novacentrix_aluminum, ure_pressurization_rate, nanoscale_aluminum, alex_aluminum, table_aluminum_particle, novacentrix_aluminum, highest_pressurization_rate, pressurization_rate_value, aluminum_particle, composite_peak_pressure

The PRA analysis of CCDC dataset is performed using the window size '8'. The results of the analysis are as shown in the Table 3. The PRA analysis has picked 3 papers from CCDC dataset which are shown in 'picked papers' section. The top 10 extracted keyphrases for these papers are shown in 'PRA keyphrases' section. The analysis shows that the number of PRA extracted keyphrases are less compared to that of LDA-RAKE analysis. The quality of the extracted keyphrases has improved. But, most of the keyphrases speak about same topics like in the paper 'Recent Developments for Prediction of Power of Aromatic and Non-Aromatic Energetic Materials along with a Novel Computer Code for Prediction of Their Power' most of the keyphrases refer to one topic 'energetic compound'. Also, there are no matches between PRA extracted papers and LDA extracted papers.

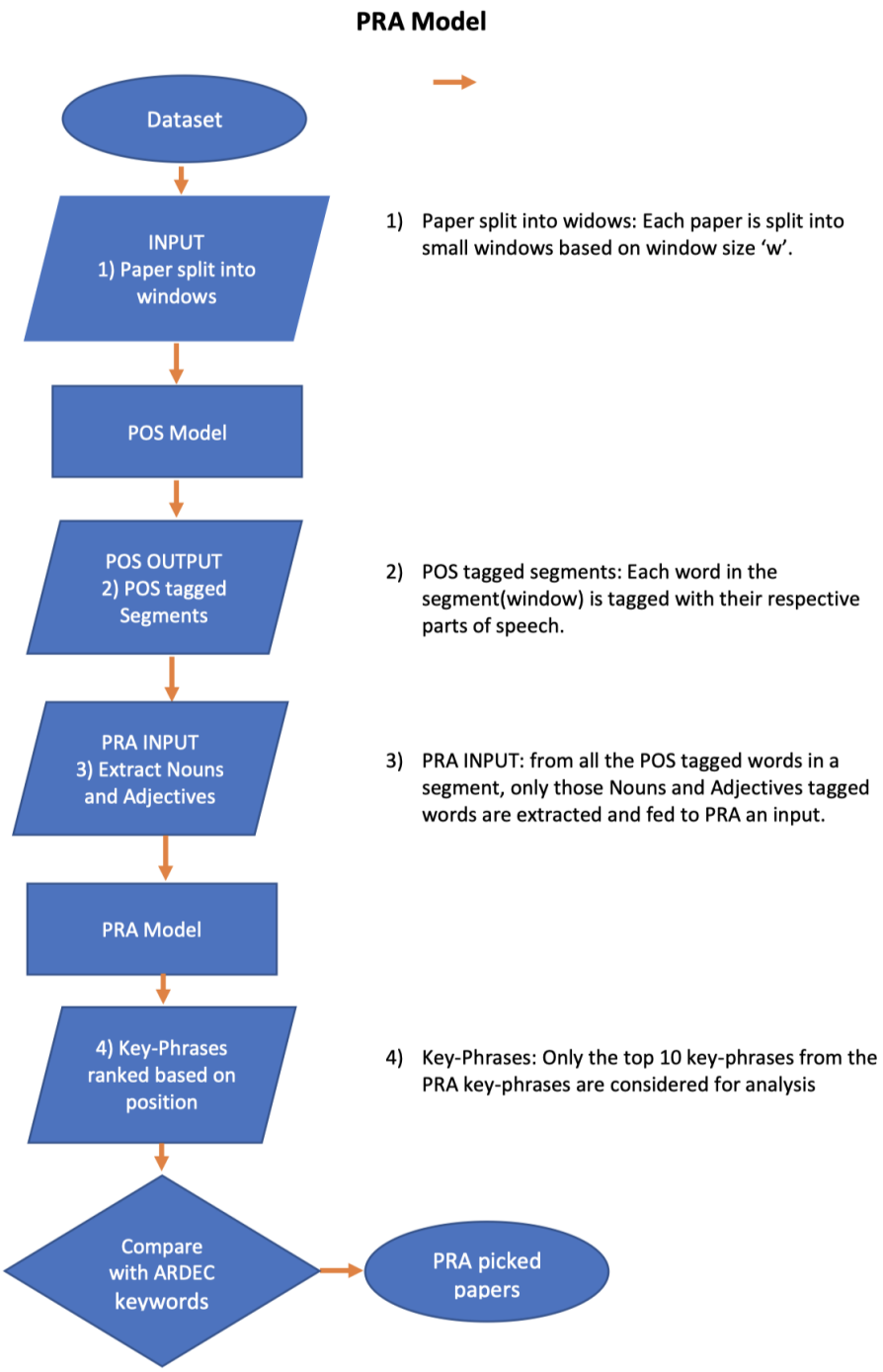

Figure 2:PRA Flow chart

4.1.1.5 LDA-PRA

We finally propose an LDA-PRA model which combines the strengths of these two approaches. PRA uses nouns and adjectives of the paper as a building block for the keyphrases.

We propose to construct a model that uses the topic keywords generated by LDA in place of the nouns in this scenario, since it is expected to have more relevance to the paper than simply nouns.

In LDA-PRA model instead of using all the nouns in the paper, the LDA model is executed first and a dictionary of topic keywords is generated for each paper and the PRA analysis is executed on only those nouns that are also picked as LDA keywords for the paper along with the adjectives. The flow chart of the LDA-PRA analysis is as shown in the Figure: 3.

As the second step POS analysis is performed, in the third step the nouns are compared with the LDA keywords. The nouns which match with the LDA keywords are given as an input for PRA analysis.

In the fifth step only the top 10 keyphrases are extracted. In the sixth step, the extracted keyphrases are compared with SME keyphrases, if at least one match is found between the extracted keyphrases and the SME keyphrases then the paper is tagged.

Ten topics are extracted for every paper in the LDA section of the LDA-PRA analysis and in the PRA analysis section of every paper in CCDC dataset is performed using the window size '8'. The results of the analysis are as shown in the Table 4. The LDA-PRA analysis has picked 4 papers from CCDC dataset which are shown in 'picked papers' section. The top 10 extracted keyphrases for these papers are shown in 'PRA keyphrases' section. The analysis shows that the number of PRA extracted keyphrases are less compared to that of LDA-RAKE analysis. The keyphrases extracted are more versatile compared to that of Only-PRA analysis. Also, there are no matches between the PRA extracted papers and LDA-PRA extracted papers, there is one match between the LDA extracted papers and LDA-PRA extracted papers which is "Multi-Parameter Study of Nanoscale TiO 2 and CeO 2 Additives in Composite AP/HTPB Solid Propellants".

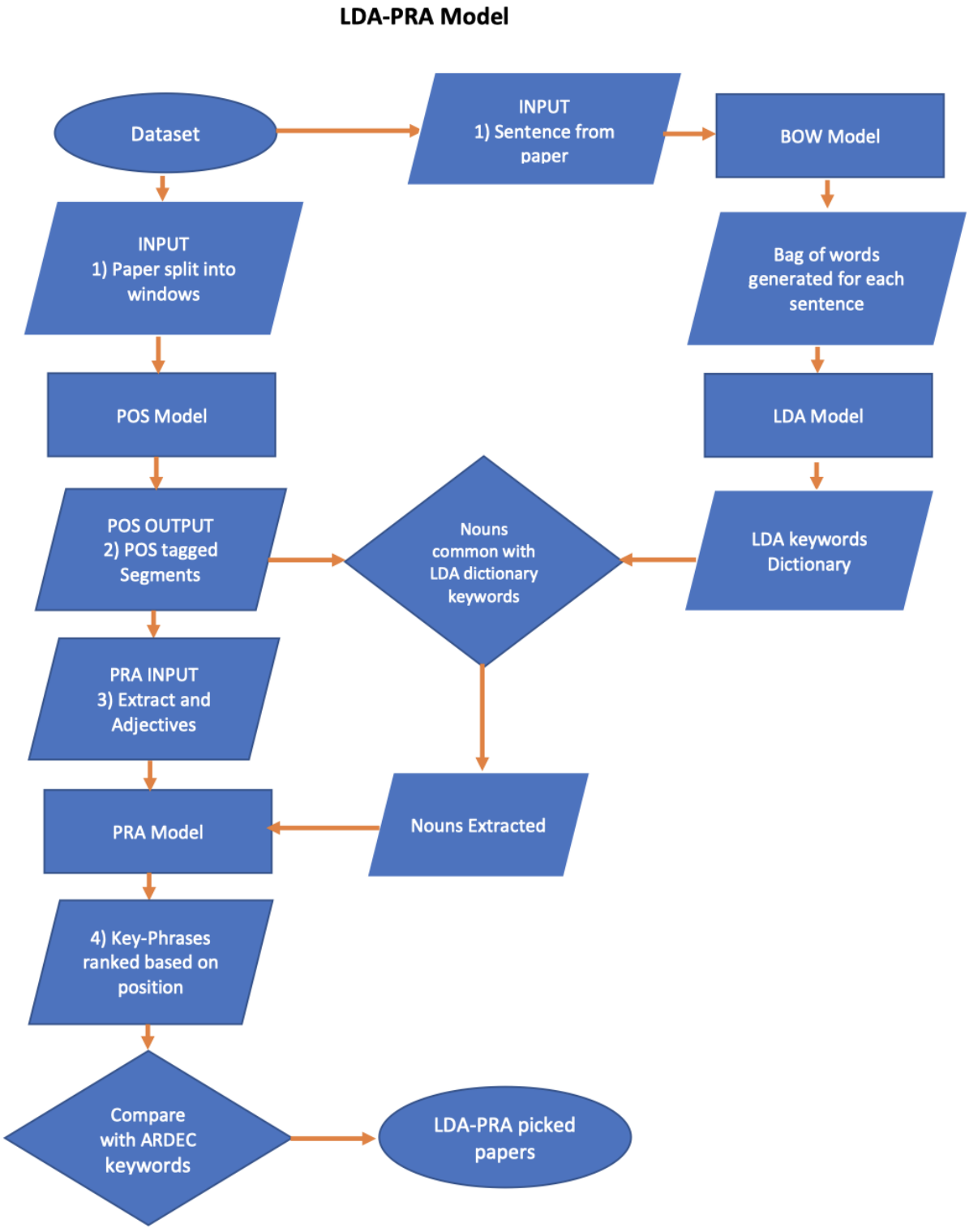

Figure 3: LDA-PRA Flow chart

Table 4: Auto-tagged papers using LDA-PRA and the corresponding keyphrases for CCDC dataset

LDA-PRA Identified papers	LDA-PRA Keyphrases
High-Energy Pollen-Like Porous Fe2O3 /Al Thermite: Synthesis and Properties	Reduced_pressure_thermites, nano_thermite, thermite_table, thermite_morphology, high_heat, nanoscale_thermite, high_energy, largest_exothermic_heat, thermite, physical_mixed_method
The Effect of Silicon Powder Characteristics on the Combustion of Silicon/Teflon/Viton Nanoenergetics	Active_silicon_content, active_silicon, silicon_teflon_viton, nanoscale_silicon, silicon_type, high_oxygen_content, maximum_spectral_intensity, ratio_silicon, teflon_viton_sitv, reactive_surface
Multi-Parameter Study of Nanoscale TiO 2 and CeO 2 Additives in Composite AP/HTPB Solid Propellants	Additive_matrix, statistical_taguchi, additive_formula, additive_effect, baseline_matrix, nanoscale_titania, pressure_sensitivity, nanoscale_ceria, formula_matrix, study_baseline
Preparation and Characterization of Nanoenergetics Based Composition B	Large_number_density, nanoenergetics, prepared_ssgt_thermal, produced_molding_powder, molding_powder, large_surface, spray_drying, explosive_consisting, ssgt_shock, nanoscale

4.1.2 SCORING METHODS:

The Precision and Recall and F-Scores have been used for calculating the efficiency of the models. Let the set of standard keyphrases be denoted by $C_{standard}$ which can either be the set of author provided keyphrases or the SME provided keyphrases. Let $C_{extract}$ denote the set of keyphrases that are extracted using the model. Then $C_{correct}(C_{standard} \cap C_{extract})$ is defined as the set of keyphrases that are common between both the $C_{standard}$ and $C_{extract}$.
The precision(p) score is defined as the ratio of number of correct keyphrases $C_{correct}$ to the number of extracted keyphrases $C_{extract.}$

$$p = \frac{c_{correct}}{c_{extract}}$$

The recall(r) score is the ratio of number of correct keyphrases $C_{correct}$ to the number of standard keyphrases $C_{standard}$.

$$r = \frac{c_{correct}}{c_{standard}}$$

Given the precision(p) and recall(r) scores the F-Score is defined as follows

$$f = \frac{2pr}{p + r}$$

These metrics will be used to score the performance of the algorithms for the various models described before.

4.1.3 RESULTS

The Datasets used in this project are described in Section 4.1.4. This section describes the results obtained for each of the models for several datasets.

4.1.3.1 ACS Dataset analysis:

The ACS Dataset is a collection of research papers from ACS Journals mentioned in described in Table 11. The PDF to text converting tool PDFMiner [6] has been used for converting all the papers from pdf to text format. The research papers of ACS dataset have different styling formats. Some papers have different heading styles and some of them are missing the heading for the introduction sections. Because of the styling formats lot of noise was added to the text output in the form of special characters, all these special characters have been removed from the original text using python filters.

This is the first dataset that has been collected for this project. The journals for this dataset are collected keeping SME keyphrases in mind. Hence only LDA-RAKE analysis was performed on this dataset. The SME keyphrases are used as the standard keyphrases $C_{standard}$ for calculating the precision, recall and F-Scores for LDA-RAKE analysis. The average values of the precision, recall and F-scores for this analysis is shown in Table 5.

Table 5: Average (over 681 papers in ACS dataset) Precision, Recall, F-Scores for LDA-RAKE based on CCDC provided keyphrases

Number of LDA topics	LDA-RAKE			
	Average Precision	Average Recall	Average F-Score	Total number of papers picked
10	0.1	0.166	0.125	19
15	0.1	0.166	0.125	19
20	0.1	0.166	0.125	13

The analysis shows that the precision, recall and F-Scores are same for all the different number of topics.

This is likely because the standard keyphrses $C_{standard}$ are taken as the same for all the 681 papers and the selected number of papers are too few to make a difference in the precision and recall score ratios. Also, since the F-Score is a factor of precision and recall the average F-score is also same for all the topics.

4.1.3.2 NUS Dataset:

The NUS Dataset is one of the standard datasets used the Position Rank Analysis (PRA) [4]. In the PRA model the analysis has been performed on only the Title and Abstract sections. But in this project the analysis has been performed on all the contents of the paper except Title and References section. The title section was not considered for the analysis because all the PRA generated top keyphrases are matching with the titles of the paper this makes the results too predictable.

Snice this dataset is not in scope of finding any CCDC keyphrases, for this analysis the author keyphrases of each paper are used as the standard keyphrases $C_{standard}$. The average scores for various number of LDA topics is shown in Table 6.

Since the number of LDA topics affects only the LDA-PRA analysis, the results of Only-PRA analysis stands same for all number of topics. The maximum number of picked papers by the LDA-PRA is obtained at 15 number of LDA topics.

Table 6: Average Scores of Precision , Recall, F-Score for Only PRA, LDA-PRA for NUS dataset

Number of LDA topics	Only-PRA				LDA-PRA			
	Average Precision	Average Recall	Average F-Score	Total number of papers with matching keyphrases to gold standard	Average Precision	Average Recall	Average F-Score	Total number of papers with matching keyphrases to gold standard
10	0.1549	0.1412	0.1377	71	0.1586	0.1498	0.1436	75
15	0.1549	0.1412	0.1377	71	0.1558	0.1469	0.1402	77
20	0.1549	0.1412	0.1377	71	0.1589	0.1489	0.1427	73

As can be seen from Table 6, LDA-PRA outperforms Only-PRA in all metrics considered.

4.1.3.3 CCDC Dataset analysis:

The CCDC dataset is a collection of papers from random publication years of "Propellants Explosives Pyrotechnics" Journal listed in Table 11. The CCDC dataset is tested on

LDA-RAKE, Only-PRA and LDA-PRA models. This analysis is performed twice by taking SME keyphrases as standard keyphrses $C_{standard}$ and author keyphrases for each paper as standard keyphrses $C_{standard.}$

The average scores of precision, recall and F-scores for Only-PRA and LDA-PRA with CCDC keyphrases as standard keyphrses $C_{standard}$ is as shown in Table 7. The results for Only-PRA model are same for all number of topics since the Only-PRA results are independent of number of LDA-topics. The precision, recall and F-scores are same for all the number of topics and models because the standard keyphrases are same for all models and very few number of papers have been picked by the models compared to that of total number of papers.

The average scores of precision, recall and F-scores for LDA-RAKE, Only-PRA and LDA-PRA with author keyphrases as standard keyphrses $C_{standard}$ is as shown in Table 8. This analysis is performed on only 139 papers out of 684 papers because the author keyphrases are available only of 139 papers.

Table 7: Average Precision, Recall, F-Scores for CCDC dataset using CCDC keyphrases as gold standard

Number of LDA topics	Only-PRA				LDA-PRA			
	Average Precision	Average Recall	Average F-Score	Total number of papers picked	Average Precision	Average Recall	Average F-Score	Total number of papers picked
10	0.1	0.166	0.125	3	0.1	0.166	0.125	4
15	0.1	0.166	0.125	3	0.1	0.166	0.125	3
20	0.1	0.166	0.125	3	0.1	0.166	0.125	4

Table 8: Average Precision, Recall, F-scores for CCDC datasets using author provided keyphrases as the gold standard.

Number of LDA topics	Only-PRA				LDA-PRA			
	Average Precision	Average Recall	Average F-Score	Total number of papers picked	Average Precision	Average Recall	Average F-Score	Total number of papers picked
10	0.1632	0.4248	0.2291	87	0.1395	0.3289	0.1914	48
15	0.1632	0.4248	0.2291	87	0.1395	0.3348	0.1925	48
20	0.1632	0.4248	0.2291	87	0.1387	0.3273	0.1909	49

As can be seen from these tables, when CCDC keyphrases are used as the gold standards, there is no measurable difference in performance. This is perhaps due to the smaller number of papers and keyphrases. Surprisingly, we find that the LDA-PRA does not do better than Only-PRA when the author provided keyphrases are used as the gold standards. We are currently working on the reasons why this may be so. It is possible that the number of LDA topics is limiting the results, since this is an artificial number. This may also be due to selecting the top n keyphrases for comparison, which was set to 10, in the above tables. The other possibility is that there is some anomaly in the dataset (perhaps too many abbreviations, etc.). These will have to be investigated in the future.

LDA-RAKE does not have a mechanism of ranking keyphrases, so unlike the Only-PRA and LDA-PRA, the keyphrases generated here do not have any limit. The following two tables list the precision, recall and F-scores for the LDA-RAKE for CCDC dataset using the author provided and CCDC provided keyphrases as the gold standards.

Table 9: Average Precision, Recall and F-Scores of LDA-RAKE analysis using CCDC keyphrses as gold standard

Number of LDA topics		LDA-RAKE			
	Average Precision	Average Recall	Average F-Score	Average number of topic keyphrases	Total number of papers picked
10	0.09	0.166	0.125	68.6666	6
15	0.09	0.166	0.125	65.1666	6
20	0.1	0.166	0.125	58.5	4

Table 10: Average Precision, Recall and F-Scores of LDA-RAKE analysis using author provided keyphrases as a gold standard

Number of LDA topics		LDA-RAKE			
	Average Precision	Average Recall	Average F-Score	Average number of topic keyphrases	Total number of papers picked
10	0.2396	0.5790	0.3308	67.36	101
15	0.2287	0.5572	0.3165	58.75	101
20	0.2239	0.5467	0.3097	54.4	96

As can be seen from this table, although the precision and

4.1.4 DATASETS AND DATA COLLECTION

Several datasets were collected and used to test the algorithms presented in this report. The first set was based off of the keywords supplied by CCDC. These are referred to as CCDC and ACS-dataset in Table 11. The third dataset is a standard dataset used by the NLP community and is called the NUS.

Table 11: Datasets Used in this Project

Dataset	Articles	Number of Papers
CCDC	Propellants_Explosives_Pyrotechnics2010, Propellants_Explosives_Pyrotechnics2011, Propellants_Explosives_Pyrotechnics2013, Propellants_Explosives_Pyrotechnics2014, Propellants_Explosives_Pyrotechnics2015, Propellants_Explosives_Pyrotechnics2016, Propellants_Explosives_Pyrotechnics2017, Propellants_Explosives_Pyrotechnics2018	684
ACS Dataset	ACS Environmental Science & Technology 2011 ACS Nano 2007 ACS Nano 2011 ACS Environmental Science & Technology 2012 70 Random papers selected from ACS journals targeting the keywords provided by CCDC	681
NUS Dataset	This is one of the original datasets that was used for Position Rank Analysis(PRA) [4]	215

The ACS Dataset: The ACS dataset was constructed based on the CCDC's interest in the journals indicated in Table 11. A vanilla search on these journals for papers with the keyphrases provided by CCDC was conducted which yielded 70 papers. We then added to this set by randomly including papers from the journals mentioned above, resulting in a total dataset size of 681 papers. The drawback of this dataset is that it does not always have author provided keyphrases. The journal papers are also of different lengths and may or may not contain distinctly identified abstracts.

The **CCDC dataset** is a random collection of publications from "Propellants Explosives Pyrotechnics" journal. This was one of the journals of interest to CCDC. This dataset is a collection of a total of 684 papers.

In order to further stress, test the algorithms, we used standard dataset used in the literature, **the NUS dataset** is one of the original datasets which is used in the Position Rank Analysis(PRA) [4], this data set is a collection of 215 papers.

4.1.5 CONCLUSION

Several methods of keyphrase extraction were compared using several metrics for the problem of identifying relevant keyphrases form scientific documents. LDA, a topic modeling mechanism was used as a component in two of these methods. We proposed two methods called LDA-RAKE and LDA-PRA that uses LDA in different ways with RAKE and PRA respectively. Extensive data collection resulted in creation of two new datasets for the journals of interest to CCDC. Experiments indicate that LDA-PRA performs well on standard datasets like NUS in terms of precision, recall and F-score. For the specialized dataset, CCDC, we find that LDA-PRA does not perform as well as only PRA. We believe this may have something to do with the specific nature of the dataset or the number of topics picked by LDA. Further experiments will be necessary to determine this for sure. As a by product of this work, we also identified a subset of papers for the SME's consideration based only on the keyphrases provided by the SME. Using this method, the workload of the SME can be cut down significantly by only having to read through a subsection of the vast data repository. For example, in some cases, the papers of interest can be whittled down to 4, from the full dataset of 684 papers.

5 BIBLIOGRAPHY

[1] D. M. Blei, A. Y. Ng and M. I. Jordan, "Latent Dirichlet Allocation," *Journal of Machine Learning Research,* vol. 3, pp. 993-1022, 2003.

[2] Ř. Radim, "gensim," 09 2018. [Online]. Available: https://radimrehurek.com/gensim/models/ldamodel.html. [Accessed 11 2018].

[3] S. Rose, D. Engel, N. Cramer and W. Cowley, "Automatic keyword extraction from individual documents," *Text Mining: Applications and Theory,* 2010.

[4] F. Cornelia and C. Corina, "PositionRank: An Unsupervised Approach to Keyphrase Extraction from Scholarly Documents," *Association for Computational Linguistics,* vol. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), p. 1105–1115, 2017.

[5] P. Lawrence, B. Sergey, M. Rajeev and W. Terry, "The PageRank Citation Ranking: Bringing Order to the Web," *Technical report, Standford Digital Library Technologies Project,* vol. Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, 29 January 1998.

[6] Manning, C. D, M. Surdeanu, J. Bauer, J. Finkel, S. J. Bethard and D. McClosky, "Stanford CoreNLP," Stanford University, 2014. [Online]. Available: https://stanfordnlp.github.io/CoreNLP/index.html.