EXAMINING THE EMOTIONAL TONE IN POLITICALLY POLARIZED SPEECHES IN INDIA: AN IN-DEPTH ANALYSIS OF TWO CONTRASTING PERSPECTIVES

DwijendraNath Dwivedi * Aravind Kumar Pandey ** and Aditya Dhar Dwivedi***

1. Introduction

The utility of topic modeling for political speeches in larger texts is quite evident in contemporary democratic process. It is a method that facilitates the identification of words that frequently appear together throughout the text and offers a comprehensive understanding of the themes present in a collection of speeches.

Topic modeling is a branch of text analytics that utilizes statistical techniques, such as probabilities, and a statistically significant data set to organize information. Not only does it aid in navigating documents, but it can also be utilized for making recommendations. Latent Dirchlet Allocation is a specific topic modeling technique that employs words in a document to generate a list of topics that are likely to occur. The model employs a probability distribution to determine the words most likely to be included in each topic. This process can be conducted both in real-time and retrospectively.

In this article, we introduce a new framework for performing topic modeling on political speeches. The framework is designed to facilitate the review of a vast number of political speeches and is a data-driven, automated approach. By applying the model to a set of documents, the computer can recognize words that frequently appear together. With this information, the analyst can then commence the analysis of the content. Despite its simplicity and effectiveness, the framework has room for improvement, particularly in regard to the selection of parameters.

2. Literature studies

In this study, Mohd Zeeshan Ansari (2020) attempted to extract political sentiments from tweets and model them as a supervised learning problem. O'Connor (2010) conducted an analysis of surveys on consumer confidence and political opinion and discovered that they correlated with the frequency of sentiment words in contemporaneous Twitter

* Krakow university of Economics, Rakowicka 27, Kraków, 31-510, Kraków, Poland.

** Aravind Kumar Pandey,Indira Gandhi National Open University India

*** Dr Ram Manohar Lohia Awadh University, Ayodhya

messages. Tumasjan (2010) used the context of the German federal election to examine if Twitter is used as a platform for political discussion and if online messages on Twitter accurately reflect offline political sentiment. Conover (2011) outlined several methods for predicting the political affiliation of Twitter users. Mishra (2016) discussed sentiment analysis of Twitter data, including the existing tools for sentiment analysis, related work, framework used, and a case study to demonstrate the methodology and results. Marchetti (2012) found a better correlation with Gallup's Presidential Job Approval polls. Wagner (2013) reported that internet usage had a positive impact on political knowledge, political participation, and attitudes towards the United States/West, but a negative effect on trust in government in Middle Eastern countries with limited government filtering practices.

In 2021, Gupta and colleagues attempted to utilize contextual analysis of text to determine the factors that influence user sentiment towards a product or service. Dwivedi and colleagues in 2020 conducted sentiment analysis and theme modeling of the government response to the COVID-19 pandemic and compared the situation in the UAE and Saudi Arabia. In 2021, Dwivedi and colleagues performed topic and sentiment analysis on Twitter data to identify concerns related to data quality and impurity. In 2022, Dwivedi and colleagues attempted to use contextual analysis of texts to categorize Twitter data regarding feelings towards COVID-19 vaccination and to highlight key concerns. The team also analyzed medical research conducted by the United Arab Emirates versus the World Health Organization to identify key themes and used text contextual analysis to categorize Twitter data based on positive and negative feelings linked to the ethical challenges of AI.

The authors Alghamdi and Alfalqi (2015) in their study found that new tools and techniques were necessary to effectively manage, search, index, and analyze large amounts of data as they witnessed the rise in electronic documents and archives. Hofmann (2001) introduced two main approaches in Natural Language Processing (NLP) and statistical methods such as thematic modeling to analyze data. Unlike NLP methods that focus on grammatical structure and parts of speech, statistical and thematic models are primarily based on the "Bag of Words" (BoW) concept. In BoW, a collection of texts is quantified in a document-term matrix, which captures the frequency of each word (columns) in each document (rows). One of the early researchers in topic modeling, Deerwester et. al. (1990), used Semantic Latent Analysis (LSA) and Singular Value Decomposition (SVD) to present one of the first topic models.

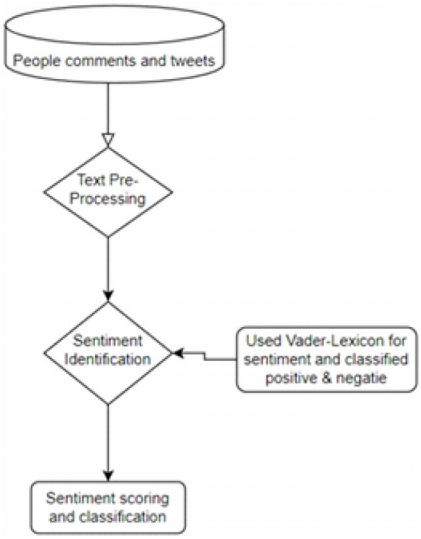
Asmussen and Moller (2019) presented a unique framework that utilized topic modeling techniques to conduct a comprehensive review of a vast collection of articles using the Latent Dirchlet Allocation (LDA) method. There are two main approaches to automatic

document processing: supervised learning and unsupervised learning. In supervised learning, a manual coding process is applied to a set of documents, which can be time-consuming. On the other hand, unsupervised learning methods like topic modeling do not require manual coding and save time in conducting an exploratory review of large collections of papers. Gotipati et al (2018) used subject modeling and data visualization to analyze student feedback from seven post-graduate courses taught at the Singapore University of Management. They compared the results of rules-based methods and statistical classifiers in extracting topics. Al-Obeidat et al. (2018) proposed a sandbox for extracting opinions and analyzing emotions to extract questions and their associated emotions from a database using LDA for theme extraction and a "Bag of Words" sentiment analysis algorithm. Polarity was determined based on the frequency of positive/negative words in the document. Benedetto and Tedeschi (2016) outlined standard approaches to social media sentiment analysis and cloud-based issues. Ajeet Ram Pathak et al. (2021) proposed a method that operates at the sentence level to extract the subject using online Latent Semantic Indexing with regulation constraints. Md. Mokhlesur Rahman et al. (2021) studied the factors related to positive and negative sentiments about reviving the economy during the COVID-19 global crisis in the United States, taking into account situational uncertainties, economic slowdown, emotional factors like depression, and related changes in work and travel patterns due to lockdown policies. Jikyung (Jeanne) Kim et al. (2022) showed that excessive consumer reactions to negative news and negative feelings intensify this excessive reaction, leading to negative consequences in livestock farming.

3. Methodology

In our study, we followed a two-step approach. Firstly, we analyzed Twitter posts and extracted positive and negative sentiments by utilizing the Naive Bayes Classifier. This probabilistic algorithm employs Bayes' Theorem to categorize the text's polarity and our analysis used 5 sentiment categories: strongly positive, positive, neutral, negative, and strongly negative. In the second step, we employed the Latent Dirchlet Allocation method to uncover the themes and recurring keywords in the text corpus. This method is commonly used to quickly and efficiently analyze a large set of polarized texts to determine the most frequently discussed topics. We also applied Latent Semantic Analysis and Singular Value Decomposition for text clustering. Clustering group's observations in a dataset based on their similarity, where words within a group are alike and the comments between groups are different. In the context of text mining, clustering divides a collection of tweets into various groups based on the presence of similar themes.

Figure 1: Figure: Process flow for topic modeling (Dwivedi. et. al. 2021)



Preprocessing text: This step is required for text analysis to transform human language in to a machine-readable format for subsequent processing and analysis. There are some mandatory steps to request clean-up, which are listed below.

- Convert all the text to lowercase
- Removing stop words, sparse terms, and particular words
- Convert numbers into words or remove the numbers
- Removing white spaces (leading and ending spaces)
- Removing punctuation (all types of special characters or symbols)

First, we have started to eliminate duplication of rows, and it is essential to delete duplicate data or rows to avoid unbiased results. Convert all text to lowercase to prevent more than one copy of the same word. For example ("drinking water" is considered to be two different words).

We were deleting punctuation because it adds more information.

Handling text data: In addition, this will shrink the size of the training dataset. We eliminate keywords that often appear in the text or we create a list of keywords, or we use predefined libraries. We used stop word and text, blob libraries that will deal with stop words . We have deleted common words in the general scenario, but we can also delete

naturally occurring comments from our textual data. We can therefore check the ten words that occur frequently, and then decide which to delete.

Spelling correction of the text: We have seen tweets with many spelling mistakes, or short words will be used. In this situation, the spell-checking step is useful to reduce the number of copies of the word. For this, we have a Text blob library it will handle spelling mistakes.

Tokenization is the process of dividing the text into a series of words or phrases. In our example, we used the text blob library to transform our tweets into blob and convert them into a group of words.

Stemming refers to the removal of suffices, like "ing," "ly," "s," etc., through a straightforward rules-based approach. For that, we will use Porter Stemmer of the NLTK library.

Lemmatization is a more suitable method than stemming because it converts the term to its root term, rather than just stripping it enough. it uses vocabulary and proceeds to a morphological analysis to obtain the root word. hence, we generally prefer to use lemmatization instead of stamping.

We've done all the basic preprocessing steps to clear the text, and now we need to extract the characteristics using natural language techniques.

N-grams are defined as a combination of several words used in combination. N-grams, bigrams, and trigrams were used. Unigrams will not have a great deal of information compared to bigrams and trigrams. We use these bigrams or trigrams to grasp the structure of the language, such as which letter or word is likely to follow that given. Those recommendations are going to depend on the implementation of our study. Sometimes, if we use low grams and do not grasp the essential differences or if we sometimes take long grams, it will not capture the overall sense of the expression.

Part-of-speech tagging (POS)

The marking of a part of the speech assigns mostly speeches to each word of the text according to its context and its definition (nouns, verbs, adjectives, and others).

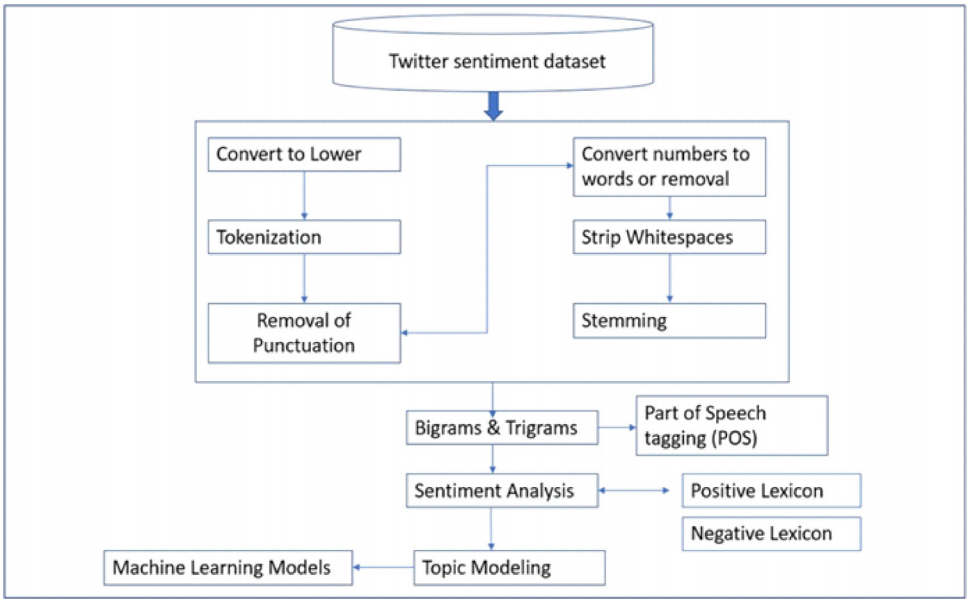


Figure 2: Figure: preprocessing process for sentiment analysis (Dwivedi et. al. 2021)

As presented in the fig 2A, firstly we started removing the duplication of rows to avoid unbiased results. Further, converted all the text into lower cases to prevent multiple copies of the same word. For example ("Crypto Currency" crypto currency" will be considered as two different words). Followed by removing punctuation as it might add any extra information or reduce the size of the training dataset while handling text data. Also, eliminated stop words that are frequently occurring words in the text by using text blob library in python. The tweets with many spelling mistakes, or short words were observed in the twitter data, hence spelling correction step are performed with the help of text blob library.

After the above steps, tokenization was done to divide the text into a sequence of words or sentences, transforming our tweets into a blob and then by converting them into a series of words. Followed by Stemming refers to the removal of suffices, like "ing," "ly," "s," etc., by a simple rule-based approach by using Porter Stemmer from the NLTK library of python. Some may use Lemmatization as it is more practical option than stemming because it converts the word into its root word, rather than just stripping the suffices. It makes use of the vocabulary and does a morphological analysis to obtain the root word. Therefore, researchers usually prefer lemmatization over stemming.

After basic preprocessing steps of cleaning the text extracted the features using the following natural language techniques. N-Grams which identify the combination of multiple

words are used together. We have used N-grams, bigrams, and trigrams. Unigrams has not captured much information as compared to bigrams and trigrams. Thus, used bigrams or trigrams to capture the language's structure, like what letter or word is likely to follow the given one. Further, part-of-speech tagging mainly assigns speeches to each word of the text based on its context and definition (nouns, verbs, adjectives, and others).

Secondly, topic modeling is the process of extracting or obtaining required features from the bag of words. This is an important technique since each word present in the corpus has considered as a feature in natural language processing. This feature reduction will help us to focus on the right content instead of going through the entire text in the training data. There are many methods used for topic modeling, Latent Dirichlet Allocation (LDA) is one of method which is used to analyze the topic modeling in the present study.

LDA is a statistical and graphical model used to obtain relationships between multiple documents in a corpus. It is developed using the variation exception maximization (VEM) algorithm for obtaining the maximum likelihood estimate from the whole corpus of text. Traditionally, this can be solved by picking out the top few words in the bag of words. However, this completely lacks the semantics in the sentence. This model follows the concept that the probabilistic distribution of topics can describe each document, and the probabilistic distribution of words can explain each topic. Thus, it helps to get a much clearer vision of how the topics are connected. It considers all corpus of entire documents in the data. After preprocessing of the corpus, each bag of words consists of common words. Using LDA model, the topics related to each document has been derived and can group all corpuses into a particular group for further usage. The flow chart below details the process of topic modeling

4. Results

A sentiment analysis engine has four stages: The first one involves breaking down the raw text into component parts known as tokens. The tokenization process analyzes the text and identifies keywords that reflect the emotions of the writer. The second stage involves isolating the sentiment-bearing tokens and disregarding the rest. The third stage involves assigning a polarity score to each of the token components. This is achieved by consulting sentiment libraries, which vary based on the program being used. For instance, in our data, we utilized the text-blob library in Python to determine the sentiment polarity of each tweet. We transformed the text data into word vectors and used the text-blob library to calculate sentiment scores for the tweets. The text-blob library employs a naive Bayes classifier to assess the polarity of a sentence. This classifier generates a score ranging

from -1 (strongly negative) to +1 (strongly positive). This is a critical aspect of sentiment analysis as it aims to understand the orientation or opinion of the sentiments expressed in the text. The orientation is quantified by a positive or negative value known as polarity. In order to classify the sentiment as positive, neutral, or negative, we analyzed the polarity scores and evaluated their distribution. We also eyeballed the sentiment expressed in the text to determine the score ranges for very positive, positive, neutral, very negative, or negative sentiment. The following is the data distribution of the sentiments in Modi's speeches.

Row Labels	Count of sentiment_label
negative	1
positive	49
Grand Total	**50**

Here is the same for Rahul Gandhi's speeches.

Row Labels	Count of sentiment_label
negative	16
positive	21
Grand Total	**37**

4.1. Topic Modeling:

The basic concept behind using traditional data mining methods in topic modeling involves converting unstructured text data into structured numerical data. The purpose of topic modeling is to identify relevant features from a collection of words, known as the "bag of words". This is crucial as each word in the corpus is considered a feature in natural language processing. By reducing the number of features, we can concentrate on the relevant information rather than having to analyze the entire text in the training data. There are several methods for topic modeling, and LDA is one of the techniques we have employed in our research. Latent Dirichlet Allocation (LDA) is a statistical and graphical model used to uncover relationships between multiple documents in a corpus. It utilizes the Variational Bayes Maximization algorithm to obtain the maximum likelihood estimate from the text corpus. The traditional approach to solving this involves selecting the top few words from the bag of words, but this fails to consider the semantics of the sentences. LDA is based on the idea that the probabilistic distribution of topics can describe each document, while the probabilistic distribution of words can describe each topic. This allows for a clearer understanding of the connections between topics. The LDA model considers the entire corpus of documents in the data, after preprocessing. The resulting bag of words consists of common words. Using this model, one can determine the topics related to each document

and group the entire corpus into relevant categories. The screenshot below shows the results of our study.

Topic	Keywords
0	forest, employment, yatra, river, karnataka, telangana, percent_commission, pocket, kerala, indian
1	kannada, extra, decreased, karnataka, yatra, group, explain, politicians, engineers, allowed
2	telangana, conversation, sitting, santosh, profession, dream, narendramodi, joker, officers, beautiful
3	affects, units, alloys, mismanagement, scared, allies, missiles, setter, payroll

Topic	Keywords
0	rajkot, jammu_kashmir, teacher, broad_gauge, teachers, jeevan_mission, water, morbi, planet, gujarat
1	steel, mahakal, khadi, himachal, sector, banaskantha, kerala, science, lothal, innovation
2	uttarakhand, kutch, jamnagar, sports, gujarat, ayodhya, chowk, border, stubble, badri_vishal
3	ayodhya, cheetahs, himachal, dairy_sector, logistics, cancer, karnataka, police, kartavya, justice

5.Conclusion

Speeches could appeal to negative or positive emotions. We can understand the same from the speeches of the two leaders from India over same time using text mining. It revealed that Narendra Modi appealed more positively in terms of positive emotions he invoked than Rahul Gandhi. Some of the key themes of Modi had been promoting Khadi, innovation, promoting tourism, promoting environmental issues and highlighting key achievements in sports. What is the key topic that Rahul Gandhi has been focusing was related to mismanagement of the government and self-praise related to "Bharat Jodo Yatra"?

When talking about political discourse, the analysis of feelings can help determine the general tone of the discourse and the emotions expressed by the speaker. It may also give an overview of the audience's reaction to the speech and the effectiveness of the speaker's message. Political speeches can often be controversial and polarizing, with different people having vastly different opinions on the same speech. Therefore, sentiment analysis of political speeches can be challenging as the sentiments expressed can vary widely depending on the listener's political affiliation, cultural background, and personal biases.

References

1. Alghamdi, R., & Alfalqi, K. (2015). "A Survey of Topic Modeling in Text Mining".

InternationalJournalofAdvancedComputerScienceandApplications,6(1),147-153.

2. Al-Obeidat et. al. (2018). "Opinions Sandbox: Turning Emotions on Topics into ActionableAnalytics".LectureNotesoftheInstituteforComputerSciences,Social- Informatics and Telecommunications Engineering, LNICST, 206, 110-119.https://doi.org/10.1007/978-3-319-67837-5_11"

3. Asmussen, C. B., & Møller, C. (2019). "Smart literature review: a practical topic modelling approach to exploratory literature review". Journal of Big Data,6(1).

4. Benedetto, F., &Tedeschi, A. (2016). "Big data sentiment analysis for brand monitoring in social media streams of cloud computing". In Studies in Computational Intelligence (Vol.639).

5. Conover M, Gonçalves B, Ratkiewicz J, Flammini A, Menczer F. (2011a) "Predict- ing the political alignment of Twitter users." In Proceedings of SocialCom/PASSAT Conference, pp.192-199

6. Deerwester,S.,Dumaiset.al(1990)."Indexingbylatentsemanticanalysis".Journal of the American Society for Information Science, 41(6), 391-407"

7. Dwivedi D.N., Anand A. (2022) "A Comparative Study of Key Themes of Scientific ResearchPostCOVID-19intheUnitedArabEmiratesandWHOUsingTextMining Ap- proach". In: Tiwari S., Trivedi M.C., Kolhe M.L., Mishra K., Singh B.K. (eds) AdvancesinDataandInformationSciences.LectureNotesinNetworksandSystems, vol 318. Springer, Singapore.https://doi.org/10.1007/978-981-16-5689-7_30"

8. Dwivedi, D. N., Mahanty, G., & Vemareddy, A. (2022). How Responsible Is AI?: Identification of Key Public Concerns Using Sentiment Analysis and Topic Modeling. International Journal of Information Retrieval Research (IJIRR), 12(1), 1-14. http://doi.org/10.4018/IJIRR.298646

9. Dwivedi, D., Vemareddy, A. (2023). Sentiment Analytics for Crypto Pre and Post Covid: Topic Modeling. In: Molla, A.R., Sharma, G., Kumar, P., Rawat, S. (eds) Distributed Computing and Intelligent Technology. ICDCIT 2023. Lecture Notes in Computer Science, vol 13776. Springer, Cham. https://doi.org/10.1007/978-3-031-24848-1_21

10. Dwivedi, D.N., Mahanty, G., Vemareddy, A. (2023). Sentiment Analysis and Topic Modeling for Identifying Key Public Concerns of Water Quality/Issues. In: Harun, S., Othman, I.K., Jamal, M.H. (eds) Proceedings of the 5th International Conference on Water Resources (ICWR) - Volume 1. Lecture Notes in Civil Engineering, vol 293. Springer, Singapore. https://doi.org/10.1007/978-981-19-5947-9_28

11. Dwivedi, D.N., Wójcik, K., Vemareddyb, A. (2022). Identification of Key Concerns and Sentiments Towards Data Quality and Data Strategy Challenges Using Sentiment Analysis and Topic Modeling. In: Jajuga, K., Dehnel, G., Walesiak, M. (eds) Modern Classification and Data Analysis. SKAD 2021. Studies in Classification, Data Analysis, and Knowledge Organization. Springer, Cham. https://doi.org/10.1007/978-3-031-10190-8_2

12. Dwivedi,D.N.,&Anand,A.(2021)."TheTextMiningofPublicPolicyDocumentsin ResponsetoCOVID-19:AComparisonoftheUnitedArabEmiratesandtheKingdom ofSaudiArabia".PublicGovernance/Zarz?dzaniePubliczne,55(1),8-22.https://d oi.org/ 10.15678/ZP.2021.55.1.02"

13. Dwivedi,DwijendraNath,etal."HowResponsibleIsAI?:IdentificationofKeyPublic Concerns Using Sentiment Analysis and Topic Modeling." IJIRR vol.12, no.1 2022: pp.1-14.http:// doi.org/10.4018/IJIRR.298646"

14. DwivediD.N.,PathakS.(2022)"SentimentAnalysisforCOVIDVaccinationsUsing Twitter: Text Clustering of Positive and Negative Sentiments". In: Hassan S.A., MohamedA.W.,AlnowibetK.A.(eds)DecisionSciencesforCOVID-19.International SeriesinOper-ationsResearch&ManagementScience,vol320.Springer,Cham.ht tps://doi.org/ 10.1007/978-3-030-87019-5_12"

15. Gotipati, S. et.al.(2018). "Text analytics approach to extract course improvement suggestionsfromstudents'feedback.ResearchandPracticeinTechnologyEnhanced Learning",13(1).

16. Gupta,A.etal.,2021."UnderstandingConsumerProductSentimentsthroughSuper- vised Models on Cloud: Pre and Post COVID". Webology, 18(1),pp.406â€"415.

17. Haselmayer, M., Jenny, M. Sentiment analysis of political communication: combining adictionaryapproachwithcrowdcoding.QualQuant51,2623-2646(2017).https:// doi.org/ 10.1007/s11135-016-0412-4

18. Hofmann,T.(2001)."UnsupervisedlearningbyprobabilisticLatentSemanticAnaly- sis. Machine Learning", 42(1-2),177-196.

19. M. M. Kim et al., "Socioeconomic factors analysis for COVID-19 US reopening sentiment with Twitter and census data," Heliyon, vol. 7, no. 2, p. e06200, 2021, doi: 10.1016/ j.heliyon.2021.e06200."

20. Marchetti-BowickM,ChambersN.(2012)"Learningformicroblogswithdistantsupervision:political

21. Mishra P, Rajnish R, Pankaj Kumar Sentiment Analysis of Twitter Data: Case Study on Digital India.INCITE-2016, Amity University, India(2016)

22. O'ConnorB,BalasubramanyanR,RoutledgeBR,SmithNA.(2010)"Fromtweetsto polls: linking text sentiment to public opinion time Series." In Proceedings of the ICWSMConference

23. sentimentanalysisof2012USPresidentialelectioncycle".,ACL(SystemDemonstra- tions) (2012), pp.115-120

24. TumasjanA,SprengerTO,SandnerPG,WelpeIM.(2010)"Predictingelectionswith Twitter:What140charactersrevealaboutpoliticalsentiment."InProceedingsofthe ICWSMConference.

25. WagnerKevin,JasonGainousDigitaluprising:TheInternetRevolutionintheMiddle East.Journal of Information Technology and Politics, 10 (3) (2013), pp.261-275

26. WangH,CanD,KazemzadehA,BarF,NarayananS"Asystemforreal-timeTwitter

Conflict of Interest:

The authors whose names are listed immediately below certify that they have NO affiliations with or involvement in any organization or entity with any ?nancial interest (such as honoraria; educational grants; participation in speakers' bureaus; membership, employment, consultancies, stock ownership, or other equity interest; and expert testimony or patent-licensing arrangements), or non-?nancial interest (such as personal or professional relationships, affiliations, knowledge or beliefs) in the subject matter or materials discussed in this manuscript.