International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

A Review on Semantic and Syntactic Similarity Measure for Political Tweets

Bijal Gadhia[1], Prapti Trivedi [2], Rashmi Hirani[3]

Submitted: 25/01/2024 **Revised:** 02/03/2024 **Accepted:** 09/03/2024

Abstract: In the modern era, social media has influenced virtually every public domain, including politics. These platforms enable users worldwide to share vast amounts of content, making social media a valuable resource for research and analysis. In countries like India, social media offers a convenient space for individuals to express their opinions on various issues, including political topics. Consequently, analyzing social media content has become a significant area of research.One key aspect of this analysis is measuring semantic and syntactic similarity within social media posts to understand user opinions effectively. This task becomes particularly challenging due to the use of informal and nonstandard language in short messages, such as tweets. Techniques like word embedding are employed to address this issue by capturing the contextual meaning of words. Additionally, factors such as word sequence and ambiguity play a critical role in deriving meaning from social media content.This review paper examines existing work related to measuring semantic and syntactic similarities in social media data. It also presents a comparative summary of various methods used for this purpose

Keywords: *Artificial Intelligence, Machine Learning, Semantic Analysis, Social Media Mining*

1. Introduction

Techniques to measure similarity between two or more part or documents of text has major application share in real world. These techniques can be used for evaluation of text answers, finding plagiarism, detect similar groups in social media etc. For text matching, meaning extraction is an important factor. There are many research work and various methods available for this. Here we have discussed some pros and cons of various methods.

Semantic Similarity

Semantic Similarity, or Semantic Textual Similarity, is a task in the area of Natural Language Processing (NLP) that scores the relationship between texts or documents using a defined metric. Semantic Similarity has various applications, such as information retrieval, text summarization, sentiment analysis, etc.[1] There have been a lot of approaches for Semantic Similarity. The most straightforward and effective method now is to use a powerful model (e.g. transformer) to encode sentences to get their embeddings and then use a similarity metric (e.g. cosine similarity) to compute their similarity score. The similarity score indicates

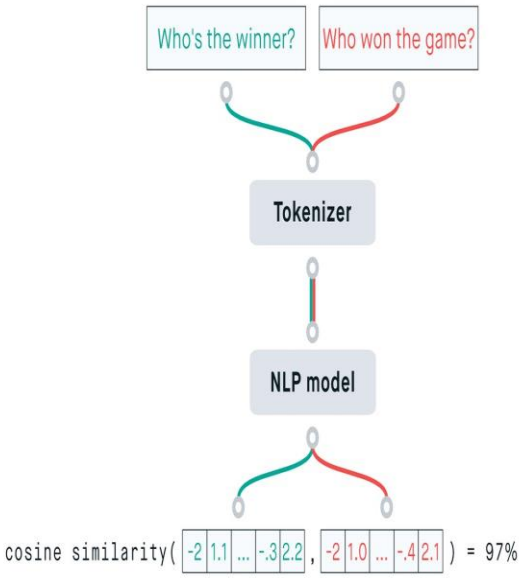

Figure 1 Tokenizer with NLP Model[1]

Whether two texts have similar or more different meanings. This post will show how to implement Semantic Similarity using Transformers, which is a powerful NLP architecture that has resulted in state-of-the-art performance for various NLP tasks. Generally it used Tokenizer and NLP Model for getting the similarity score.

1 Assistant Professor, Computer Engineering Department, Government Engineering College, Gandhinagar
2 Assistant Professor Department of Information Technology ,Dharmsinh Desai University
3 Assistant Professor, Computer Engineering Department,Silver Oak College of Engineering
** Corresponding Author Email: bij.1988@gmail.com*

Syntactic Similarity:

Natural Language Processing (NLP), also known as computational linguistics, is one such technology that is garnering the interest of many scientific researchers due to its right blend of language, machine learning, and artificial intelligence. After a detailed discussion about the use of transformer architecture in NLP in a past blog, Lumenci shares an analysis of two methods to calculate syntactic similarity in text, namely Jaccard similarity and Cosine similarity.[12] The syntactic similarity is based on the assumption that the similarity between the two texts is proportional to the number of identical words in them (appropriate measures can be adopted here to ensure that the method does not become biased towards the text with a larger word count

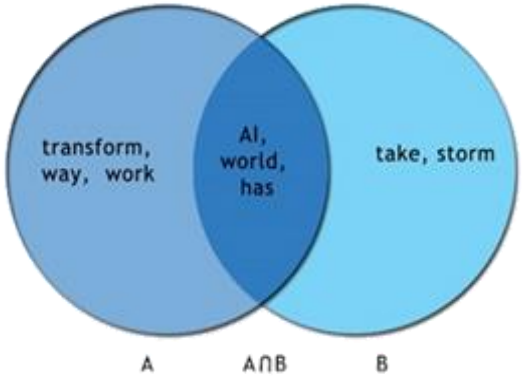

Figure 2 Syntactic Similarity Model[12]

Ontology

Ontology is an explicit specification of a conceptualization. The term is borrowed from philosophy, where Ontology is a systematic account of Existence. For AI systems, what "exists" is that which can be represented. When the knowledge of a domain is represented in a declarative formalism, the set of objects that can be represented is called the universe of discourse. This set of objects, and the describable relationships among them, are reflected in the representational vocabulary with which a knowledge-based program represents knowledge. Thus, in the context of AI, we can describe the ontology of a program by defining a set of representational terms. In such ontology, definitions associate the names of entities in the universe of discourse (e.g., classes, relations, functions, or other objects) with human-readable text describing what the names mean, and formal axioms that constrain the interpretation and well-formed use of these terms. Formally, an ontology is the statement of a logical theory [15]

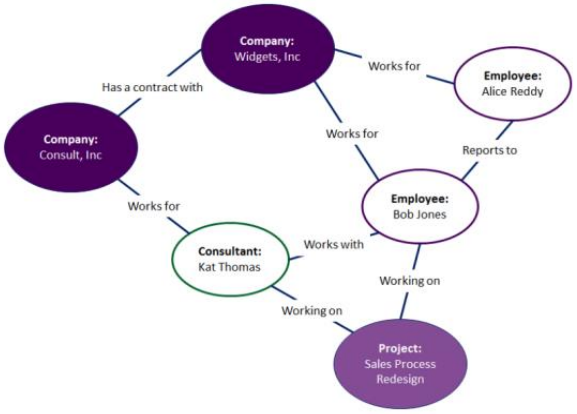

Figure 3. Example of Ontology[12]

2. Literature Review

We have reviewed some the research work done in recent time related to Semantic and syntactic measures in political domain.

In research work [1] they have applied Semantic and Syntactic Similarity Measure (TSSSM). The approach uses word embedding's to determine semantic similarity and extracts syntactic features to overcome the limitations of current measures which may miss identical sequences of words. In their work tweet Dataset is used.

They have not performed sentiment analysis which may be a key point for political tweets. The research work is also missing political term weightage and descriptive meaning (non-dictionary word) e.g. BJP, NDA, UPA etc.

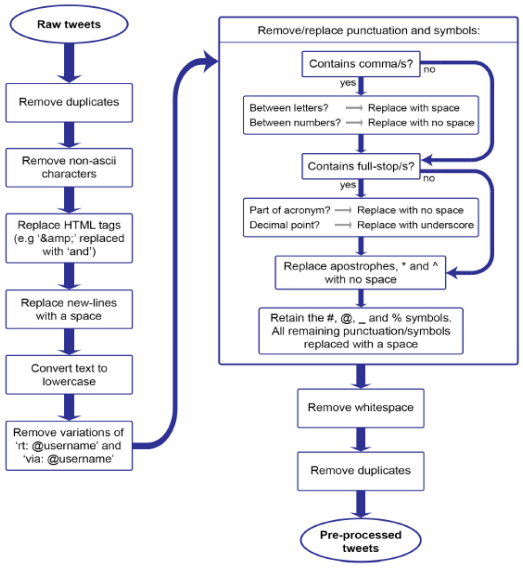

Figure 4 TSSSM Model[1]

In another research work[2], F. A. Wenando et.al., have used Unigram, Bigram, Trigram, N-Gram (1-2) and N-

Gram (1-3) and SMO Algorithm. They have also used Twitter Dataset. They have classified tweets after preprocessing in Multi class Classifications. As a result they have achieved 82.7% overall accuracy. Following table shows various algorithms and their result.

Algorithm	Accuracy	Precision	Recall
Naive Bayes	80.20	80.20	80.2
SMO	82.70	82.70	82.7
Logistic	77.70	78.60	77.1
KNN	78.80	78.60	79.1
Decision Tree	77.30	77.00	78.3

Table 1. Result Comparison in Research[2]

In research[3] M. Rodríguez-Ibáñez et. Al., have applied Semantic and Syntactic Similarity Measure (TSSSM)[1] and SKIP Gram Model. In their work two fold analyses is done. The work contains statistical and non linear features. They have also used twitter dataset and achieved 73% accuracy. In their work statistical characterization is done using indices derived from well-known temporal and information metrics and methods including entropy, mutual information, and the compounded aggregated positivity index allowing the estimation of changes in the density function of sentiment data. Feature extraction from nonlinear intrinsic patterns in terms of manifold learning using auto encoders and stochastic embedding's applied.

In research work[4], Voting classifier(LR-SGD) with TF-IDF is applied. They have tested seven machine learning techniques in their work. They have also applied their research over Twitter Dataset. Among their testing techniques, they got highest overall accuracy is 79%.

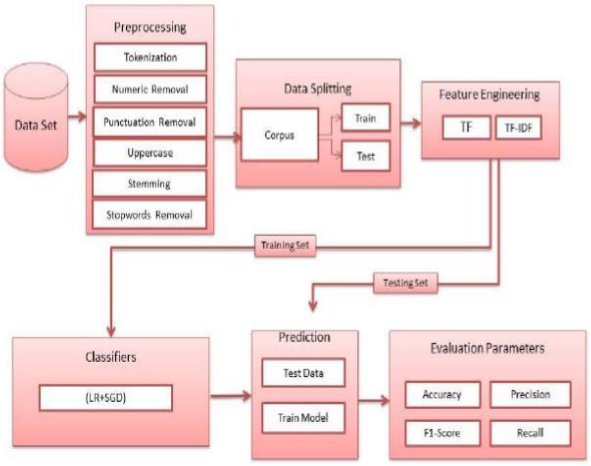

Figure 5. Model for LRSGD[4]

We have also review a research work by A. Bilbao-Jayo and A. Almeida[5]. They have applied CNN algorithm for classification of text. This categorization scheme is widely used by political scientist. In their research work they have achieved 70% Accuracy. Neural Network method CNN is applied with RELU.

Following table contains comparative summary of the research works those we have reviewed.

Ref. No.	Technique Used	Dataset	Points to Consider	Result
[11]	Semantic and Syntactic Similarity Measure (TSSSM SKIP Gram	Twitter	Uses word embedding to determine semantic similarity	0.75 MSE
[2]	Unigram, Bigram, Trigram, N-Gram (1-2) and N-Gram (1-3 SMO Algorithm	Twitter	Classified Tweets after Pre-processing	82.7 % Accuracy
[3]	Semantic and Syntactic Similarity Measure (TSSSM SKIP Gram	Twitter	Two Fold Analysis Statistical Non Linear Features	73% Accuracy
[4]	Voting classifier(LR-SGD) with TF-IDF	Twitter	Seven ML Techniques are tested.	79% Accuracy
[5]	CNN for Classification	Twitter	Categorisation scheme widely used by political scientist is applied	69.99% Accuracy
[6]	word2vec	Twitter	Unsupervised machine learning and deep neural Network model are used.	76.22% Accuracy
[7]	CNN LSTM	20NewsGroup dataset	HMM Algorithm with WEM is used.	88.4 % Accuracy
[8]	Decision Tree and K-Neighbors	Twitter	Psycholinguist categories (e.g. immigrants, security,	76% Accuracy

			kindness, etc.) are analyzed	
[9]	Python tools and libraries, TextBLOB	Twitter	User wise (Political Leader wise) Tweets analysis is done	77% Neutral Tweets
[10]	Pre-trained ML models of HAR Glove, Word2vec	Wikipedia as Word Modelling	Apply Label for Human Activity Recognition	73.3% Accuracy

3. Conclusion:

After reviewing various research work and methods related to semantic and syntactic measures from social media content, we can summarize that various features and their combination are used to extract the meaning of the content and methods like SVM, Naïve Bayes, CNN can be applied over it. Unigram, Bi-Gram, N-gram have proven their better performance.

4. Future Work:

As a future work, one can apply ontology based approach for domain specific work to get proper meaning and relation for non-dictionary words. Also sentiment analysis with this work can lead to better classification.

The authors declare no conflicts of interest.

References

[1] C. Little, D. Mclean, K. Crockett and B. Edmonds, "A Semantic and Syntactic Similarity Measure for Political Tweets," in IEEE Access, vol. 8, pp. 154095-154113, 2020, doi: 10.1109/ACCESS.2020.3017797.

[2] F. A. Wenando, R. Hayami, Bakaruddin and A. Y. Novermahakim, "Tweet Sentiment Analysis for 2019 Indonesia Presidential Election Results using Various Classification Algorithms," 2020 1st International Conference on Information Technology, Advanced Mechanical and Electrical Engineering (ICITAMEE), 2020, pp. 279-282, doi: 10.1109/ICITAMEE50454.2020.9398513.

[3] M. Rodríguez-Ibáñez, F. -J. Gimeno-Blanes, P. M. Cuenca-Jiménez, C. Soguero-Ruiz and J. L. Rojo-Álvarez, "Sentiment Analysis of Political Tweets From the 2019 Spanish Elections," in IEEE Access, vol. 9, pp. 101847-101862, 2021, doi: 10.1109/ACCESS.2021.3097492.

[4] A. Yousaf et al., "Emotion Recognition by Textual Tweets Classification Using Voting Classifier (LR-SGD)," in IEEE Access, vol. 9, pp. 6286-6295, 2021, doi: 10.1109/ACCESS.2020.3047831.

[5] A. Bilbao-Jayo and A. Almeida, "Improving Political Discourse Analysis on Twitter With Context Analysis," in IEEE Access, vol. 9, pp. 104846-104863, 2021, doi: 10.1109/ACCESS.2021.3099093.

[6] Z. Tasnim, S. Ahmed, A. Rahman, J. F. Sorna and M. Rahman, "Political Ideology Prediction from Bengali Text Using Word Embedding Models," 2021 International Conference on Emerging Smart Computing and Informatics (ESCI), 2021, pp. 724-727, doi: 10.1109/ESCI50559.2021.9396875.

[7] R. Nassif and M. W. Fahkr, "Supervised Topic Modeling Using Word Embedding with Machine Learning Techniques," 2019 International Conference on Advances in the Emerging Computing Technologies (AECT), 2020, pp. 1-6, doi: 10.1109/AECT47998.2020.9194177.

[8] M. Furini and M. Montangero, "On Predicting the Success of Political Tweets Using Psycho-Linguistic Categories," 2019 28th International Conference on Computer Communication and Networks (ICCCN), 2019, pp. 1-6, doi: 10.1109/ICCCN.2019.8847055.

[9] Y. Malhan and S. Singal, "Sentiment Analysis of Ayodhya Verdict using Twitter," 2020 International Conference on Intelligent Engineering and Management (ICIEM), 2020, pp. 504-509, doi: 10.1109/ICIEM48762.2020.9160017.

[10] K. Shimoda, A. Taya and Y. Tobe, "Combining Public Machine Learning Models by Using Word Embedding for Human Activity Recognition," 2021 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events (PerCom Workshops), 2021, pp. 2-7, doi: 10.1109/PerComWorkshops51409.2021.9431141.

[11] Bridget T. McInnes, Ted Pedersen,Evaluating measures of semantic similarity and relatedness to disambiguate terms in biomedical text, Journal of Biomedical Informatics, Volume 46, Issue 6, 2013, Pages 1116-1124, ISSN 1532-0464, https://doi.org/10.1016/j.jbi.2013.08.008.

[12] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. 2017. Graph attention networks. arXiv preprint

arXiv:1710.10903 (2017).

[13] Prashanth Vijayaraghavan, Soroush Vosoughi, and Deb Roy. 2017. Twitter demographic classification using deep multi-modal multi-task learning. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). 478–483. http://dspace.lpu.in:8080/jspui/bitstream/123456789/3422/1/11312700_5_2_2015% 202_46_40% 20PM _full%20report.pdf

[14] Wongthontham, Pornpit & Abu-Salih, Bilal. (2018). Ontology-based Approach for Identifying the Credibility Domain in Social Big Data. Journal of Organizational Computing and Electronic Commerce. 28. 354-377.

[15] "Dataset",https://www.kaggle.com/codesagar/indian-political-tweets-2019-feb-to-may-sample