**RESEARCH PAPER:**

# Predictive analytics using Social Big Data and machine learning

## LITERATURE REVIEW

TEAM SSY

Prepared by: **SRUJAN BOMMENA**

January 19, 2025

# SUMMARY AND ANALYSIS

## BACKGROUND/MOTIVATION:

The research paper discusses about how digitization of human interaction together with the growing infiltration of social media platforms, e-commerce, and digital communication channels, causes social big data(SBD) to emerge as the largest data platform. Such data include insights contained in tweets, blogs, reviews, and comments but are unstructured, noisy, and heterogeneous.

Some organizations are facing difficulties in the management, analysis, and extraction of actionable insights from this data. Current statistical techniques and models are unable to deal with the large volume of heterogeneous data.

- Twitter generates over 500 million tweets daily, revealing user sentiment, preferences, and behavior.
- Integration of data from various platforms is challenging due to user traces left on various platforms.
- Real-time processing of data is necessary for timely decision-making.
- SBD presents opportunities for increased customer engagement, market understanding, and improved decision-making, but current tools pose significant barriers.

## PROBLEMS IN EXISTING LITERATURE:

The current methods used in sentiment analysis (SBD) revolve around statistical models and the tradition of text mining algorithms like LDA and LSA. However, those models have several major shortcomings: stubbornness on topic changes, lack of temporal analysis, inability to perform well for short texts, and inability to catch nuances in language. These issues hinder organizations from effectively leveraging SBD in profit-driven tasks such as customer segmentation, sentiment analysis, and personalized recommendations. In addition, the lower semantic capability of these models yields poor insight capabilities.

**IMPORTANCE OF QUESTION IN RESEARCH PAPER:**

This study discusses the combination of Semantic Web technologies and machine learning algorithms in attempts to solve the challenges posed by Social Business Data, or SBD. It aims to improve decision-making and benefits from predictive accuracy, filling a literature gap, and enlarging prospects to various domains. This would empower businesses and researchers in analyzing user behavior, preventing misinformation and spam, and fueling innovation in data analytics.

The way second cities operate hinges on the fact that in depicting user engagement and preferences, models which predict a personalized way may grow to optimize satisfaction of a customer. We stress hybrid approaches in machine learning, together with semantic technologies, and ontology-based models.

# METHODS

**MODELS USED IN RESEARCH ANALYSIS :**

1. Logistic Regression

Logistic regression defines a statistical method for binary classifications like spam versus non-spam.

2. Naive Bayes

Naïve Bayes is classically used for spam detection and sentiment analysis.

3. Deep Learning

Deep Learning is a very potent technique that captures complex, nonlinear relations within the data and is applied to large datasets.

4. Decision Tree and Forest

Decision Trees and Random Forest provide interpretability and high-performance.

**NOTE WORHTY TECHNIQUES:** Ontology-based models help improve data understanding and structure while feature engineering provides meaningful features to improve model accuracy, and temporal analysis tracks evolving user interests.

# SIGNIFICANCE OF WORK

**KEY FINDINGS AND CONTRIBUTIONS:**

This study presents a defined architecture for predictive analytics in social big data (SBD) that underpin on machine learning, semantic web technologies, and big data framework. The framework guides a stepwise process for handling SBD, involving data generation, collection, storage, analysis, and visualization. It bridges a wide gap of statistical analysis through semantic enrichment for better data structuring and analytic precision. Decision tree models showed better performance in terms of accuracy and AUC than other tested algorithms. The role of semantic enrichment with ontologies to infer high-level topics from short terms would cater to certain limitations of existing methods. The integration of WordNet, DBpedia, and IBM Watson NLU enabled advanced knowledge inference, improving predictive outcomes. The study provides practical insights for organizations to leverage on predictive analytics.

**RESULTS:**

The results achieved at the end of the paper are;

1. Integrates unstructured data with structured datasets for informed decision-making.
2. Overcomes limitations of traditional approaches with ontology-driven semantic enrichment.
3. Captures trends and changes in user behavior over time for improved predictive power.
4. Uses big data technologies like Hadoop/MapReduce for efficient, real-time data processing.
5. Framework is not restricted to one domain and can be customized for various applications.

**IMPLICATIONS FOR FUTURE PRACTICE:**

Integrating semantic technologies with predictive analytics is opening an applied, scalable frame for research, linking theoretical developments on machine learning with practical applications in big data analytics. The research, therefore, has the potential to spur innovation in social media data analytics by way of improving spam detection, customer segmentation, and real-time sentiment analysis. Such flexibility in the frame invites predictive analytics into other domains, notably health, education, e-commerce, and politics. Future directions include real-time analytics improvement by introducing mobile cloud computing with IoT for real-time analytics, employing fuzzy logic to tackle data uncertainty, and developing multi-domain ontologies that allow simultaneous classification and prediction in various fields.

# CONNECTION TO OTHER WORK

**RELATE THE PAPER TO OTHER RELEVANT STUDIES:**

Traditional machine learning models were used for semantic analysis in some previous studies, like IBM Watson NLU. This model addresses shortcomings of LDA and LSA, in modeling topics and in producing results lacking semantic depth. The proposed model-relatives incorporate interpolations that rely on ontology interoperability and domain knowledge extraction.

**HOW DOES THIS PAPER BUILD ON OR DIFFER FROM PREVIOUS WORK?**

This paper concerns the ontology-based semantic enrichment for text mining, augmenting topic modeling and circumventing some limitation of traditional statistical models. It uses WordNet, DBpedia, and IBM Watson NLU for disambiguation of semi-structured corpuses. Baseline predictive models such as Gradient Boosted Trees, Random Forests, and Deep Learning are utilized for better prediction accuracy and scalability. Temporal segmentation for the evolution of user behavior and interest is also introduced. The study is domain-specific and concerns political sentiment analysis and

customer segmentation. The framework integrates into its design, end-to-end coupling with external knowledge bases, and advance methods of evaluation.

**CITATIONS**:


- Blei, D.M., Ng, A.Y., and Jordan, M.I. (2003). "*Latent Dirichlet Allocation*." Journal of Machine Learning Research.
- Friedman, J.H. (2001). "Greedy function approximation: A gradient boosting machine." Annals of Statistics.


# RELEVANCE

### REFERENCES TO MY CASPTONE PROJECT

My project explores machine learning and predictive analysis, incorporating the tools, methods, and evaluation techniques outlined in a relevant research paper. The primary focus of the project is on interpreting and analyzing data.

### IDENTIFY ANY SPECIFIC METHODS, THEORIES, OR FINDINGS THAT YOU MIGHT INCORPORATE INTO YOUR PROJECT.

1. Predictive Modeling.
2. Machine Learning.
3. Model Evaluation Techniques.
4. Tools mentioned in the research such as Scikit learn, Pandas/Numpy, deep learning analysis.

**HIGHLIGHT ANY POTENTIAL AREAS WHERE YOUR CAPSTONE COULD EXPAND UPON OR DIVERGE FROM THE PAPER'S FINDINGS.**

The proposed capstone project intends to showcase the ability of predictive analytics to incorporate multi-domain classification, real-time analytics, deep learning models, data imbalance, explainable AI, mobile cloud and IoT integration, ethical and privacy considerations, and custom visualization tools. However, the project may also explore unique methods to achieve cross-domain integration, custom visualization tools for particular use cases, or different frameworks for addressing data imbalance.