# Jordan University of Science and Technology

## Paraphrase identification and semantic text similarity analysis in Arabic news tweets using lexical, syntactic, and semantic features

**Authors:** Mohammad AL-Smadi, Zain Jaradat, Mahmoud AL-Ayyoub, Yaser Jararweh

**Abstract:** The rapid growth in digital information has raised considerable challenges in particular when it comes to automated content analysis. Social media such as twitter share a lot of its users? information about their events, opinions, personalities, etc. Paraphrase Identifica- tion (PI) is concerned with recognizing whether two texts have the same/similar meaning, whereas the Semantic Text Similarity (STS) is concerned with the degree of that similarity. This research proposes a state-of-the-art approach for paraphrase identification and se- mantic text similarity analysis in Arabic news tweets. The approach adopts several phases of text processing, features extraction and text classification. Lexical, syntactic, and seman- tic features are extracted to overcome the weakness and limitations of the current tech- nologies in solving these tasks for the Arabic language. Maximum Entropy (MaxEnt) and Support Vector Regression (SVR) classifiers are trained using these features and are evalu- ated using a dataset prepared for this research. The experimentation results show that the approach achieves good results in comparison to the baseline results.