ITU KALEIDOSCOPE
ONLINE2020

7-11 December 2020

A TECHNIQUE FOR EXTRACTING THE INTENTION OF MESSENGERS IN SOCIAL MEDIA

Balakrishnan Nalin Prashanth
University of Moratuwa, Sri Lanka

Session: 4

Digital transformation in daily life

Paper S4.1

Overview

- Purpose of the research
- Similar work
- Proposed technique
- Results and achievements
- Conclusion and acknowledgements

Why do we need to identify the Intention of Messengers in Social Media?

- To protect from fraud
- Especially, to safeguard children from the ones who misuse social media
  - Solicitation of sex, promoting violence/terrorism, theft of sensitive information, bullying, etc.

- Why social media?
  - Social media has become an integral part of daily life
  - Almost all social media support Instant Messaging (IM)
  - IM via social media protects the anonymity of the users

Similar Work

- “Structural analysis of chat messages for topic detection” (H. Dong, S.C. Hui, and Y. He, 2006)
  - An indicative term-based approach
  - Support Vector Machine (SVM) classifier using set of topic indicative terms
- “Topic detection in instant messages“ (H. Zhang, C.D. Wang, and J.H. Lai, 2014)
  - Identify topic using message-word co-occurrence matrix information
- “Intention extraction from text messages” (I. Song, and J. Diederich, 2010)
  - Segments messages to sentences, convert the sentences to tuples using dialog act classifier

Proposed Technique

- Understand text message classification
- Identify short-comings of existing approaches and add value
- Identified values
  - Richness of meaningful content
  - Importance of meaning provided via non-text components
  - Meaning hidden due to language complexities and human error (misspelled words)
  - Identifying related messages and semantically grouping them together

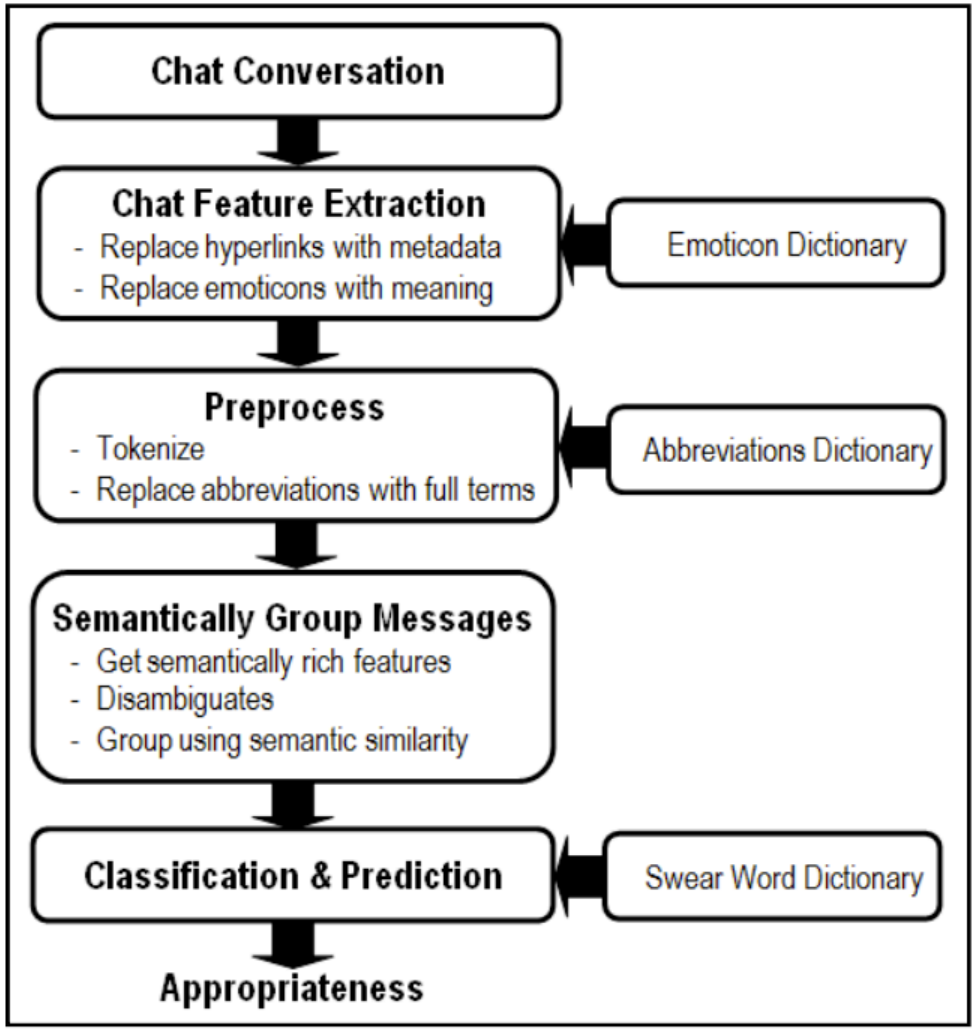

Figure 1: high-level logic

Proposed Technique

- Extracting text from selected non-text components
  - Hyperlinks – meta-data description
  - Emoticons – using an emoticon Unicode dictionary
- Handling language complexities and human error
  - Abbreviations – using an abbreviations dictionary
  - Misspelled words or short words – using a phonological spell-checker along with a disambiguation module
- Grouping semantically related messages
  - Identify the best matching synset (Wordnet)
  - Group identified synset lists (different messages) using a novel algorithm
- Identify or classify the grouped messages
  - Identify swear words in input to classifier
  - Support Vector Machine (SVM) based classifier

Experiments, Results and Achievements

Algorithm 1 Similarity Algorithm

```
similarityScore = 0
FOR synset A in synsetList A DO
    FOR synset B in synsetList B DO
        try:
            similarityScore = similarityScore + similarity (synset A, synset B)
        except TypeError
            try:
                similarityScore = similarityScore + similarity (synset B, synset A)
            except TypeError
                similarityScore = similarityScore + 0.0
    END FOR
END FOR
similarity = 4*(similarityScore / (len(synsetList A)*len(sysetList B)))
return similarity
```

Figure 2: Similarity algorithm

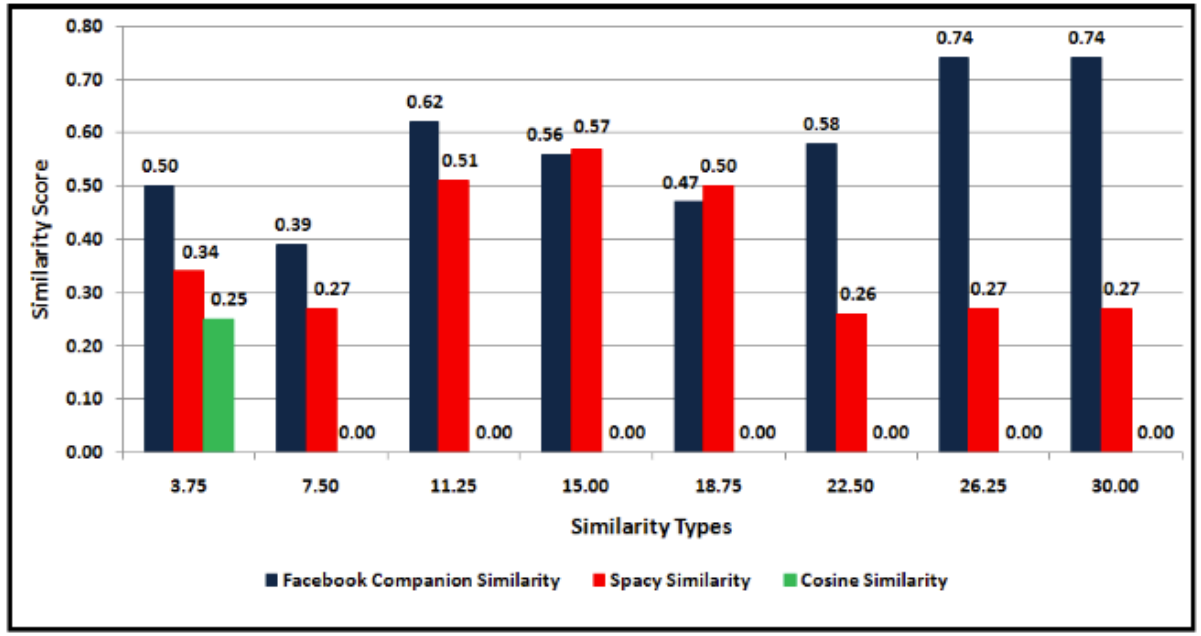

Figure 3: Similarity scores experimented

Experiments, Results and Achievements

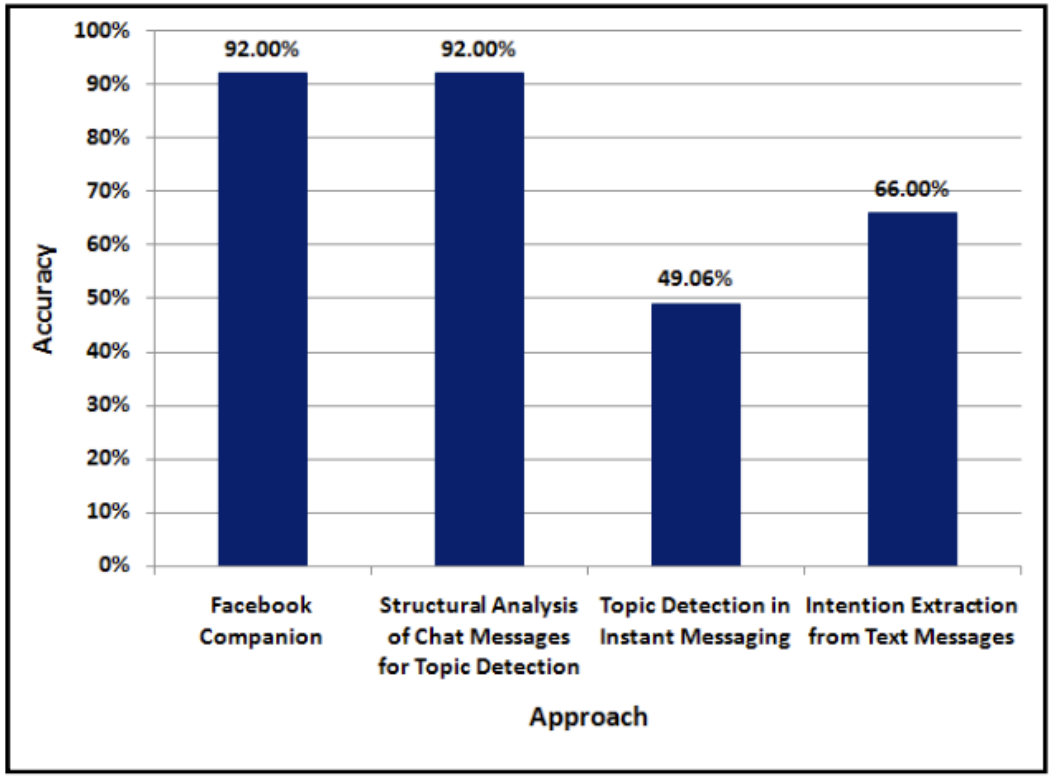

Figure 4: Overall accuract comparisons

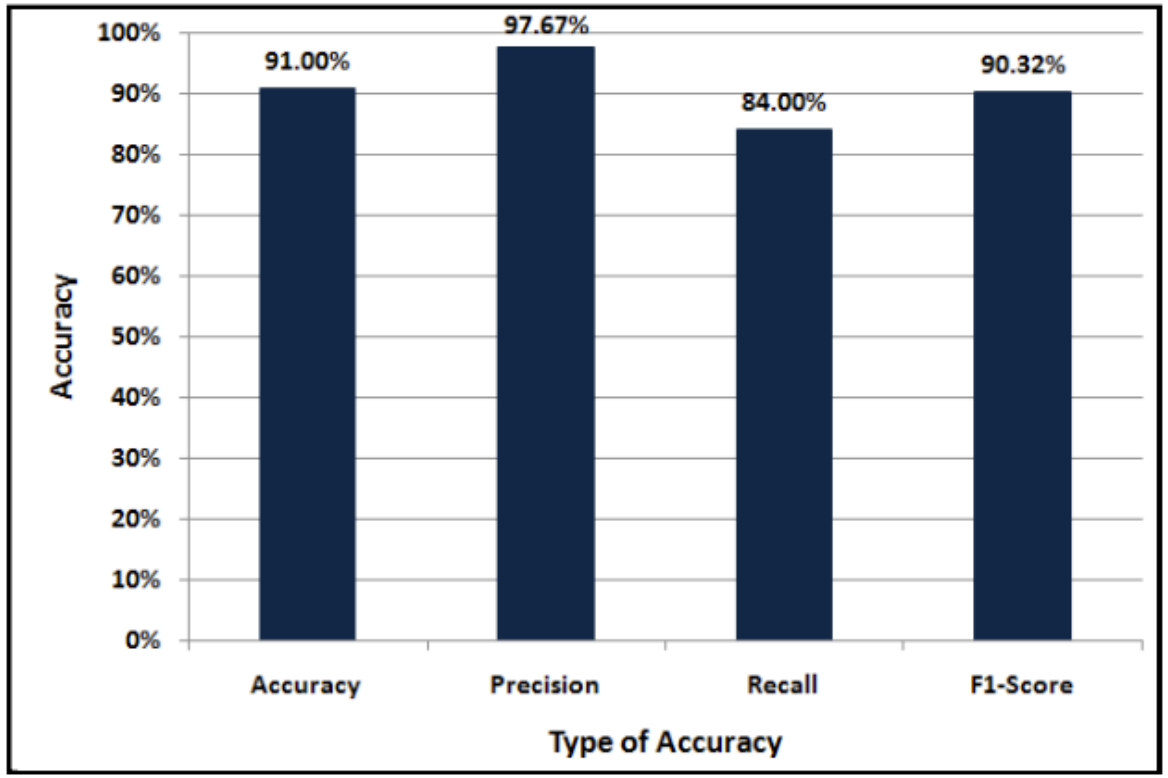

Figure 5: Accuracy types

Conclusion and Acknowledgements

- Conclusion
  - Algorithm to group similar text messages
  - Important to enrich content which contribute to add meaning to text
- Future work
  - Multi-language support
  - Improvements to the spell-checker
  - Dictionaries can be dynamic and growing
  - Support graphics
  - Expand the classification criteria
- Major contributions and findings
  - Overall need to enrich text to classify instant messages
  - Techniques used for extracting textual meaning from selected non-text components
  - Technique used to identify and handle misspelled words
  - Similarity Algorithm
- Acknowledgements
  - Dr. Mohamed Firdhous
  - All the related contributions and all those who supported

ITU KALEIDOSCOPE ONLINE2020

Thank you!