Dude, srsly?: The Surprisingly Formal Nature of Twitter's Language

Yuheng Hu, Kartik Talamadupula, Subbarao Kambhampati

Name: FNU, Arpana Hosabettu, **NetId** : fa97

In the paper , 'Dude,srsly?:The Surprisingly Formal Nature of Twitter's Language', the authors conduct a study on the characteristics of the language used on twitter in comparison with other medium like SMS/blogs/newspaper. In this paper authors aim at providing a formalized method to measure Twitter's language in terms of linguistic and psycholinguistic features as the author feels that previous research on twitter have not focused on quantifying the language and differentiating it from other medium.

Twitter is a micro blogging site and it is logical for its content and language to be different compared to other mediums. Not only did the authors give a very compelling justification of this conventional notion, but they also quantified the twitters language itself in terms of linguistic and psycholinguistic styles. The study was indeed exploratory in recognizing the general set of aspects and quantifying them. Although the mathematical equations in SOCLIN was not described in detail, this provides a way to infer psycholinguistic properties in sparse documents sets.

However, there are lot of aspects like geography, language, and gender that are not considered in the study. Geography and language of the user may influence the language used in media. The study makes a mention of first and third pronoun being dominant in twitter. One of the study shows that women tend to use first pronoun heavily than men. All these aspects may have huge impact on the comparison numbers provided. Also, emoticons, hashtags play important role in social media which when considered may lead to very interesting results.

The datasets also have drawbacks. The kind of data collected does not specify the randomness of the data, geographical distribution of its users and the gender, social relationships and network. For instance although 45million tweets have been collected, they are geographical distributed whereas email collected is between employees which is a closed set. As the author themselves mention, the time periods of the data do not match and may have influenced the results. Larger sets of data may also provide a very different result.

The paper compares twitter quantitatively with each of the individual mediums like SMS, IM, blogs, email, newspaper which is not intuitive to the reader in deciding the formalness the language. This approach fails to convey a general idea of Twitter's language being more formal. One of the approach that could be employed is to aggregate the mediums in terms of informal conversation based medium like SMS/IM and a formal information sharing based mediums like emails/newspaper. This would provide the some probabilistic data on the style of twitters language.

In summary, although this study is revealing, it is not conclusive. There is still scope for further study using larger and relevant data sets, identifying larger aspects that may indirectly influence the language like geography and gender, uncommon aspects of a particular media like hashtags, emoticons.