Social Identity and Computational Linguistics: A Review

Sunil K Sonawane, Avinash B. Surnar

***Abstract*— Language Varies as Social Nature of people varies, as it tends to change with demographic variations. It can be seen that people use language differently depending on their social status, economic status, Gender, Age etc. In India caste groups tend to use language differently than other groups and have both structural and phonological noticeable features."Evolution in language is Phenomenon of these variations in language over demographic groups of people". In this paper our aim is to provide overview of Computational research in Demographic themes about Linguistics with focus on language and social identity of people (Gender, age, geographical area etc.)**

***Index Terms*— Age, Sociolinguistics, Author, Prediction**

I. INTRODUCTION

Internet has changed the model of scientific research. More digital data is available, resulting data based Research has more works than traditional approaches like theory development, narrating of natural phenomena and discovery. This data driven approach has given birth to Computational Linguistics (CL), it is computational study to extracts Information from language and evaluate structure of verbal communication. This relation between language and social variation is mutually dependent as Using language specifically represents social identity of a person and people use language specifically to represent themselves in society. Because of this dynamics, social change is connected to language use. For example females use some features in language more often than their males.

"Sociolinguistics is study of effect of language and social variation on each other."Traditional approaches for sociolinguistics used systematic studies of people and cultures, surveys datasets created were small in size and they were formed to perform statistical and manual observations on the data. The tremendous data available on internet in the form of social media in the form of posts, bogs, email, chats, online presence of news and other textual data gave sociolinguists to try their hands on more broad study of social dependence of language (and lingual dependence of social groups), they must now select appropriate tools to manage this tremendous data and analyze, process it, with traditional approaches it was not possible, so they partnered themselves with Field called CL. Using methodologies of CL, Sociolinguists can achieve what they are trying their hands on. Sociolinguistics can help us build better NLP tools, NLP is vaster field, CL and sociolinguistics can help us develop improved NLP models. Theories in sociolinguistics can help CL community to build foundations for their research.

Computers made research in social science simple and the term computational social science came into existence. Increased interest in building models for sociality of language and analyzing social aspect of language combined fields of CL and Sociolinguistics often termed as "Computational Sociolinguistics" named together [1].Researchers in this new field are Sociolinguistics who need methodologies of CL and CL researchers who need theory from Sociolinguistics.CL uses its own techniques but it is multidiscipline research practice.

As considering scope of this discussion, CMC (computer mediated Communication) can be used as data and techniques of CL can be applied to spoken language and its social aspects. Relation between Verbal and non verbal communication can be analyzed when they are occurring at the same time.HCI (Human Computer Interaction) can be improved using computational sociolinguistics. Computational Stylometry can be studied in deep which can be used for applications as author profiling, authorship attribution and plagiarism detection taking styles of the writers of text in consideration. Study on features other than lexical and stylistic variation can be explored, other social variables than age, gender and location can be studied, languages like Marathi, Hindi, Tamil, Telugu etc can be explored using Computational sociolinguistics. video and audio data can be used as basis of study in computational sociolinguistics.

II. METHODS OF COMPUTATIONAL SOCIOLINGUISTICS RESEARCH

Communities of sociolinguistics and CL are collaborating with each other, though they have goals of parent community, these communities of researchers' influences the work in counterpart area. Hence, Sociolinguistics and other social science fields are helping build more effective CL models and CL is helping sociolinguistics by creating tools and redefining the theoretical models.

Mostly work in past few decades on CL is based on development of new approaches to computational modeling. Some examples are neural networks, probabilistic graphical

***Sunil Sonawane**, Department of Computer Science and Information Technology, Dr. Babasaheb Ambedkar Marathwada University, Aurangabad, India, +91 8552082329.*
***Avinash Surnar,** Department of Computer Science and Information Technology, Dr. Babasaheb Ambedkar Marathwada Universisty, Aurangabad, India, +91 9921615326.*

All Rights Reserved © 2018 IJARCET

models and deep learning approaches. There continuous renovation in the model building because of requirement for greater prediction accuracy when there is output as correct answer. There is overlap between modeling approaches with CL and sociolinguistics. For example, logistic regression is widely employed by variationist sociolinguists using a program called VARBRUL (Tagliamonte 2006). Similarly, logistic regression is widely used in the CL community, especially in combination with regularization methods when dealing with thousands of variables, for example for age prediction (Nguyen et al. 2016).

While CL focuses on technical aspects like Prediction and creativity, Sociolinguistics focuses on reliability and validity of the model. Thus, when techniques from CL are adopted in sociolinguistics, Validation approach is considered a priority.((Krippendorff 2013).Bagging have been widely used approach for many models for Computational Sociolinguistics. It can be said that CL is quantitative and Sociolinguistics is qualitative field.

In the field of sociolinguistics, Interviews, surveys and observations were sources of data in ancient times, these were labor based and time consuming methods. As these types of datasets were small in size they were rarely used in CL. Manual methods in sociolinguistics were used to study these datasets. When CMC emerged digital data was available in tremendous amount with low cost and less efforts. (Doyle 2014). Some examples of the Internet data are Micro blogs, web forums and online review sites (Johannsen et al. 2015).There are also difficulties challenges and limitations of these data sources, They are prone to error, social media do not represent whole population, additional data correction and processing is needed using parsers and entity recognizers etc. Data privacy and suitability for work also plays a role in data collection for data collection.

III. Language and Social Uniqueness

We will discuss about modeling language variation associated with social variables like age, gender, location etc. using computational point of view in this section. Authors use language in a way that represents their social identity y (Bucholtz and Hall 2005).language, dialect or styles are chosen by the author to shape their communication. knowingly or unknowingly speaker adjust language delivery to represent themselves in a social circumstances. (Wardhaugh 2011).

Knowing that language varies with social context, many studies are done on extracting the social context from language like identifying age and gender. These automatic systems for author's age, gender or social variable prediction are proved to be more useful and accurate than human based prediction (Burger et al. 2011; Nguyen et al. 2013).although some studies are based on predicting other variables like ethnicity, and social class, huge work is dedicated to age, gender and geographic location prediction. Facts like difference in cultural activities like marriage for age in male and female and variation in language is more often in younger people can be basis for study on age and sex dependent social variables.

Earlier studies were based on corpus data like British national corpus, recorded data, blogs etc. recent studies use more social media data like twitter. Other internet sources like LinkedIn, Facebook, YouTube, IMDB, email have also been studied. Fig 3.1 shows a generalized concept of how prediction of sociolinguistic variable is done using computational approach. As shown in fig 3.1, Internet based text like social media posts and blogs are crawled and collected as raw data, this data is processed to get useful data suitable for classification and non useful elements like smiley's, special symbols, url links are removed from data. We cannot feed data as it is to classifier, we need to extract features from the data that can be used to train our classifier, these features are styles, pattern that make the basis for the classification of data." It is to be noted that training data here is the labeled" with class it belongs to. Here, classes are social variables like age (male and female).The trained classifier then can be given new data and it will predict which class the author of the text belongs to.

Most studies are focused on parts of this model rather than overall prediction of social variable. Developing predictable features like character based n grams, bag of words which is more useful with combinations of character based n grams, POS features is given good thought by researchers. This resulted in development of tools like POS taggers, Linguistic Inquiry and Word Count (LIWC) (Pennebaker, Francis, and Booth (2001)).

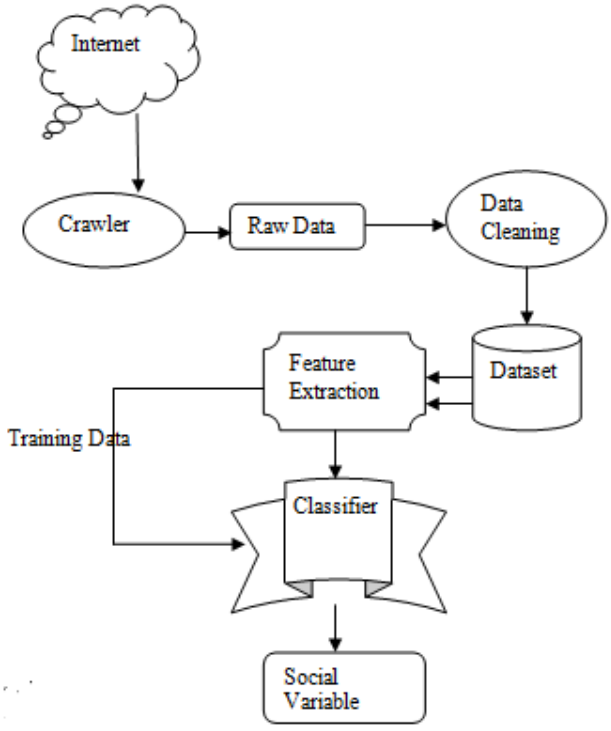

Fig 3.1 Prediction of social variable from language.

Evaluation of the labeled features supported theories in Sociolinguistic theories like male tend to use more numbers, technical words and URLs than females. (Bamman, Eisenstein, and Schnoebelen 2015),While females ten to use more family and relationship words. Grammatical features are studied in many cases POS frequencies and POS patterns

are good features to study sociolinguistics. Stylistic features are also extracted in various works, certain styles like emoticons and social media words like "omg" and "lol" are more often used by females. Genre is used as feature for few works, women work in factious and romantic writings than men.(Gianfortoni, Adamson, and Rosé (2011))

Age based features are unigrams and POS perform well. Younger people use less standard language, they tend to use words like "like" more often :D, more abbreviations and emoticons. Tasks such as sentiment analysis and topic classification can be improved using socialistic features like age and gender (Hovy (2015)).

We can adopt work and models from Computational Sociolinguistics to other fields like education, literature and health communication (Mayfield et al. 2014).Computational sociolinguistics need to be applied at multiple variables in combination. We need to develop models that can be used in multiple social contexts as previous studies are having domain specific models. We need to use more theories from sociolinguistics and social science for developing technical methodologies in CL and NLP. We need better nlp tools and more generalized models for language analysis, preprocessing tools that will make data suitable for detecting language variation.

IV. CONCLUSION

Language is not just a form of communication but it is a strong social identity and it varies with variation in social attributes. Language is used to build, represent and nurture social relationships. Thus; some aspects of the language can be predicted using computational approach. It can be concluded that text is a good source for studying aspects of human behavior and social science from works on social media and other data. In this paper we overlooked research strategies in computational linguistics in social context, we reviewed previous works in field of computational sociolinguistics, and we stated what has been achieved liable in previous works and scope that is available for possibilities, which is huge.

REFERENCES

[1]. Anders Johannsen et al. "*Cross-lingual syntactic variation over age and gender*" Proceedings of the 19th Conference on Computational Language Learning, pages 103–112,Beijing, China, July 30-31, 2015.
[2]. Dong Nguyen , A. Seza Doğruöz "*Computational Sociolinguistics: A Survey"* MIT Press Cambridge, MA, USA, pages 537-593, Volume 42 Issue 3,September 2016
[3]. Tagliamonte, Sali A. 2006. *"Analysing sociolinguistic variation"* Cambridge University Press A. Mukherjee and B. Liu., Improving Gender Classification of Blog Authors. In Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, 2010.
[4]. Krippendorff, Klaus, 2013. *"Content Analysis: An Introduction to Its Methodology"*, chapter Validity. SAGE Publications.
[5]. Bucholtz, Mary and Kira Hall. 2005. "*Identity and interaction: A sociocultural linguistic approach*." Discourse studies, 7(4-5):585–614
[6]. Wardhaugh, Ronald. 2011. "*An Introduction to Sociolinguistics*." Wiley-Blackwell.
[7]. Burger, John D., John Henderson, George Kim, and Guido Zarrella. 2011. "*Discriminating gender on Twitter*." In Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, pages 1301–1309, Edinburgh, Scotland, UK.
[8]. Bamman, David and Noah A. Smith. 2015. "*Contextualized sarcasm detection on Twitter."* In Proceedings of the Ninth International AAAI Conference on Web and Social Media, pages 574–577, Oxford, UK.
[9]. Gianfortoni, Philip, David Adamson, and Carolyn P. Rosé. 2011. "*Modeling of stylistic variation in social media with stretchy patterns.*" In Proceedings of the First Workshop on Algorithms and Resources for Modelling of Dialects and Language Varieties, pages 49–59, Edinburgh, Scotland.
[10]. Hovy, Dirk, Anders Johannsen, and Anders Søgaard. 2015. "*User review sites as a resource for large-scale sociolinguistic studies"*. In Proceedings of the 24th International Conference on World Wide Web (WWW '15), pages 452–461, Florence, Italy
[11]. Mayfield, Elijah, M. Barton Laws, Ira B . Wilson, and Carolyn P. Rosé. 2014. "*Automating annotation of information-giving for analysis of clinical conversation"*. Journal of the American Medical Informatics Association, 21(1):122–128.

All Rights Reserved © 2018 IJARCET