Ain Shams University

Faculty of Al-Alsun

English Department

A Computational Study of Males' and Females' Patterns of Language Use in Arabic and English on Twitter in 2012 – 2013

MA Thesis

Submitted by

Safinaz Muhammed Saeed Tawfiek

Under the supervision of

Dr. Khaled Elghamry

Associate professor of Linguistics

Faculty of Al-Alsun

Ain-Shams University

Dr. Nihal Nagi Sarhan

Associate professor of Linguistics

Faculty of Al-Alsun

Ain-Shams University

2020

Ain Shams University

Faculty of Al-Alsun

English Department

A Computational Study of Males' and Females' Patterns of Language Use in Arabic and English on Twitter in 2012 – 2013

MA Thesis

Submitted by

Safinaz Muhammed Saeed Tawfiek

Under the supervision of

Dr. Khaled Elghamry

Associate professor of Linguistics

Faculty of Al-Alsun

Ain-Shams University

Dr. Nihal Nagi Sarhan

Associate professor of Linguistics

Faculty of Al-Alsun

Ain-Shams University

2020

Ain Shams University

Faculty of Al-Alsun

English Department

A Computational Study of Males' and Females' Patterns of Language Use in Arabic and English on Twitter in 2012 – 2013

MA Thesis

Student Name: Safinaz Muhammed Saeed Tawfiek

Degree: MA (English linguistics)

Section: English Section

Faculty: Faculty of Al-Alsun

University: Aim-Shams University

Graduation Year: 2007

Registration Year: 2014

Acknowledgment

"Praise be to God, who hath guided us to this: never could we have found guidance, had it not been for the guidance of God". (The Glorious Qur'an: 7: 43).

Primarily, I would like to express my sincere thanks to my supervisors. Dr. Khaled Elghamry, Associate Professor of linguistics, Ain Shams University. Dr. Khaled's mentorship is a credit and honor. I am indebted and totally obliged to him. Dr. Khaled supported me not only by providing research guidelines and advice, but also academically through the bumpy road to finishing this thesis. I also extend my sincere gratitude to Dr. Nihal who has a positive influence on my development through her role as my research co-supervisor. Dr. Nihal provided me with academic guidance and supported me through my research phases.

Further, I would like to express my special thanks and gratitude to all the staff of the English Department, Faculty of Al-Alsun, where I was graduated and did my postgraduate studies. I also thank all doctors in English Department and the head of the department Dr. Samar Abdel Salam for their sincerity, enthusiasm, and commitment. Besides, I feel to acknowledge my deep sense of gratitude to Dr. Fadwa Kamal Abdel Rahman, for her support and encouragement.

Finally, I must express my very profound gratitude to my parents, my sister and my husband for providing me with unstinting support and continuous encouragement throughout my years of study and through the research process and writing this thesis. Special thanks also to my friends Mona Gamal, Sarah Abdul Hameed and Soad Mohammed Naguib for their constant inspiration. Besides, I am grateful to all of you. I would like to thank my friend Heba Magdy Fawzi, the one who supported me and helped me a lot in performing the statistical results of this study. This accomplishment would not have been possible without you.

Thank you

Safinaz Muhammed Saeed

Abstract

This study investigates the linguistic lexical choices made by 500 Egyptian Twitter users (250 males and 250 females) writing in MSA and ECA in a selected corpus of 30,000 tweets over the period 2012 to 2013. The study examines the validity of gender-based variations in computer-mediated discourse, and how this can help in authorship studies. Users are identified as males or females according to their names, alias or bio. Certain gender-preferential features, used in previous sociolinguistic and computational studies (e.g. the use of function words, words that denote insults, taboo words, intensifiers, interrogatives, etc.) are selected and applied to tweets. The research examines selected morphological, stylometric and sociolinguistic gender-based features. Perl programming language and bag of words (BoWs) model are used in running codes and representing documents as sets of words. Finally, statistical analysis is performed. On the morphological level, results show that the addition of ta ta'aneeth (the gender inflectional-suffix) to derived nouns and adjectives is a significant feature that characterizes female authors. On the stylometric level, it is revealed that the repetitive use of pronouns marks females' style, while the recurrent use of demonstratives and prepositions marks males' style. On the sociolinguistic level, results demonstrate that women tend to use insults and interrogatives more frequently, whereas males make recurrent use of taboo words and intensifiers more than females. Concerning authors' choice of domains, results highlight that females prefer to talk about their bodies and life partners, while males prefer to discuss issues related to sports, economy, and politics, in addition to using more loanwords.

Keywords: Gender, Twitter, Arabic, BoWs model, Perl programming language, Morphological features, Stylometric features, Socioloinguistic features.

Transcription

Table (1): The Pronunciation of the Letters of the Arabic Alphabet in IPA characters:

ا	/ a / / a: /	ب	/ b /
ت	/ t /	ث	/ θ /
ج	/ g /	ح	/ ħ /
خ	/ x /	د	/ d /
ذ	/ ð /	ر	/ r /
ز	/ z /	س	/ s /
ش	/ ʃ /	ص	/ ş /
ض	/ ḍ /	ط	/ ţ /
ظ	/ ẓ /	ع	/ ʕ /
غ	/ ɣ /	ف	/ f /
ق	/ q /	ك	/ k /
ل	/ l /	م	/ m /
ن	/ n /	ه	/ h /
و	/ w / / u / / u: /	ي	/ y / / i / / i: /
ء	/ʔ /	ة	/ a/ / at/

Abbreviations

ANLP	Arabic Natural Language Processing
BoWs	Bag of Words
CA	Classical Arabic
CMC	Computer Mediated Communication
ECA	Egyptian Colloquial Arabic
HS	Highly Significant Statistically
LIWC	Linguistic Inquiry and Word Count
MSA	Modern Standard Arabic
NLP	Natural Language Processing
NS	Non-Significant Statistically
P	Probability Value
Perl	Practical Extraction and Report Language
S	Statistically Significant
TF	Term Frequency
TM	Text Mining

List of Tables

List of Figures

Contents

Chapter One

With the rapid growth of online social media in recent years, analyzing the language of its users has attracted the attention of researchers. The analysis of online texts that investigates authors' language is called Authorship Analysis. Authorship Analysis has its roots in stylometry, which provides statistical analysis for some textual features. First attempts to quantify the writing style of authors go back to the 19th century with the study of Mendenhall on Shakespeare's plays on 1887. Later, Mosteller's and Wallace's study "The Federalist Papers" on 1964 is considered the most influential work in this field. "Their method is based on statistical analysis of the frequencies of a small set of common words (e.g. and, to, etc.) which produced significant discrimination results" **(Stamatatos, 2006, p. 823)**. Authorship Analysis provides means to examine a document to glean information about its author, such as: gender, age, educational background, ethnic background, etc. For this purpose, Authorship Analysis framework is adopted, in the context of studying and investigating the relationship between gender and language use.

The present thesis examines a corpus of 30,000 tweets from Twitter. Twitter is considered as synonymous to a micro-blog in the sense that its users post their tweets that consist of 140 characters or less. Twitter is chosen as a corpus for many reasons. It is an attractive site for research because of its large volume and diverse users. Unlike other social sites, the majority of its content is explicitly public. Moreover, **Ott (2016)** states that Twitter allows its never-ending stream of language to be easily

accessed and used, where anyone who knows a little about programming can download many public tweets. Besides, the language used is informal similar to everyday conversations, expressing people's different thoughts and feelings. Twitter is also selected because of "its specific characteristics based on brief communication, which differ from other social network sites such as Facebook and Myspace" **(Alvarez and Munoz, 2012, p. 38)**. Moreover, **Ugheoke (2014)** believes that Twitter has influenced the way in which business is conducted and political orientation is constructed. It has played a major role in organizing and coordinating events like the Arab spring and the Egyptian revolution in particular. In addition, **Pak and Paroubek (2010)** believe that Twitter is a reliable corpus for many reasons. According to them, Twitter is "used by different people to express their opinion about different topics…… Twitter contains an enormous number of text posts…….. Twitter audience varies from regular users to celebrities, company representatives, politicians, and even country presidents." **(2010, p. 1320).** The previously-mentioned reasons led researchers to build corpora of twitter data. They use the content of tweets to identify and study real-world phenomena, such as age and gender. Many scholars used twitter-based corpora in gender studies in both Arabic (e.g. **Hussein, et.al. (2019)**) and English (e.g. **Burger et.al. (2011)**, **Deitrick, et.al. (2012)**, **Ugheoke (2014)** and **Ott (2016**)). Twitter, in addition, is used as a corpus in other studies and proved to have significant results, e.g. O'Connor's (2010) study on the effect of Twitter on public opinion, Bollena's et. al. (2010) study on predicting future performance of the stock market, and Agarwal's et. al. (2011) study on sentiments expressed in tweets.