A Genetic-Bayesian Short Message Service Spam Filter with Text Normalization and Semantic Indexing

Erhiri, J[#1], Adebayo, A.O*[2], Akinsanya, A.O[#3], Sodiya, A.S[#4], Eze, M.O*[5], Ebiesuwa Seun[#6]
[#1*2#3#4*5#6]*Faculty, Computer Science Department, Babcock University, Ilishan-Remo, Ogun State, Nigeria*

Abstract

Ever since the first Short Message Service (SMS) service was introduced in 1993, its popularity has continued to soar over the years such that SMS communication now constitutes a major segment in the spectrum of telecommunication. The popularity and extensive usage has attracted the interest of many researchers to the inherent potential in harvesting data and metadata from collection of SMS corpus for the performance of linguistic, diachronic, normalization and sociolinguistic studies and also in the validation and comparison of different classifiers in SMS spam filters. However, freely available dataset where this type of information can be found for research purposes are quite difficult to obtain. This is mostly due to the confidentiality of SMS where users want to reveal as little of the contents of their phones as possible. This paper is geared towards the examination of the techniques adopted in the creation of SMS corpus and the ethical consideration involved in the protection of users' interest and privacy. For a successful SMS corpus creation, a main consideration is the requirement to protect the rights and interests of the message donors and any other person mentioned in the text messages, without altering the original text in order to gather sufficient metadata information. A review of existing work in the field was done to ascertain ethical observations adopted. Participant consent, data anonymization, and ensuring participants' safe information storage are basic ethical consideration adopted to ensure a successful SMS corpus creation.

Keywords: *Corpora, Corpus, Metadata, Linguistic, Diachronic, Normalization, Sociolinguistic*

I. INTRODUCTION

Short Message Service (SMS) or simply called text message is the text communication service component of phone, web or mobile communication systems, using standardized communications protocols that allow the exchange of short text messages between fixed line or mobile phone devices. The portability and ubiquity of services provided by these devices has entrenched the mobile phone and its services to be a part of our lives. The popularization of mobile phones coupled with their low cost of sending messages through these devices has lead text messages to become the most used means of electronic communication in the world today (Facebook Developer Conference, 2016).

An estimated 8.3 trillion SMS was sent worldwide in the same year2014 alone (Portio Research Report,2014). Statistics equally showed that 3.39 billion SMS was sent and received in Nigeria alone in the year 2013 (Iiloani, 2015). The tremendous rise in the usage of SMS stems from the ease of use, ubiquity in nature, high open rates, low cost of transaction and inherent trust in the channel(Cloudmark Report, 2013). With soaring degree of societal penetration, SMS has attracted the interest of researches in variousfieldsof studies ranging from linguistic, diachronic, normalization,sociolinguistic studies and usability research such as the impact of SMS on social culture, text-entry improvement, named entity recognition, authorship identification and spam message detection (Chen andKan, 2011).Despite the need forSMS dataset for research purposes, the availability of SMS corpus is in short fall. The major reason for the scarcity of SMS corpusis attributed to the private nature of SMS.

SMS are majorly stored in the mobile network operator's database due to limited storage capacity of user's devices.For legal and confidentiality reasons, network operators are not permitted to release users SMS in their custody for research, as users' messages are always personal and confidential.

Even the collection of SMS messages from individual phone by researchers, requires the observation of privacy of user. To get phone users to willingly part with their messages for the creation of SMS corpora, assurance of the protection of rights and interests of the message donors and any other person mentioned in the text messages should be guaranteed. This research therefore examines different techniques in the creation of SMS corpora and the ethical consideration of user in the creation of SMS corpora.

II. SMS COLLECTION

The major reason for the scarcity of dataset is the challenge of collecting messages from participants. Different collection measures have been adopted to cushion these challenges. Most adopted measures are dependent upon the research focus(Tagg,2009).

SMS is viewed as a personal communication between two participants, which may contain very delicate information such as bank and email details. Owners of text messages are very cautious in revealing the content of their phone for whatsoever reason; be it academic or research purposes.

Another obstacle in the collection of SMS from users is the means of extracting messages from user's phone. Mobile phone operates under many platforms and extracting messages requires the development of different application to interface these phones. Installing application to extract messages from large number of participants can be a herculean task. Apart from development of different extraction apps, phone users are unwilling to surrender their phone to strange researchers.

SMS Collectionwhich refers to the gathering of messages directly from user's phone for the main purpose of research.It starts with the recruitment of willing participants. Recruitment techniques include advertising through national media and online advertising forums such as social network sites, email lists. The use of family members and friends and professional acquaintance has yielded better results (Fairon andPaumier, 2006; Sanders, 2012; Taggs, 2009; Song et al, 2012).In order to motivate SMS donors and contributors, incentives such as cash gift, raffle draws and other promotional gift items are pledged at each level of the project(Verheijen and Stoop ,2016; Treurniet, De Clercq, Heuvel and Oostdijk,2015).

SMS collection methods can generally be categorized into three techniques. The following represents these techniques:

(i) Transcription: This is the conventional method of simply typing donated text messages from phone to a form on designated website for submission (How and Kan, 2005) or into a word processing application (Masinyana, 2008; Tagg, 2009; Elizondo, 2011). Manually copying messages from a phone with a pen to paper can equally be useful.

(ii) SMS exportentails the use of software to upload SMS from user phone to a web forms or to researcher's phone. Software suites such as Treo Desktop (supporting PalmOS) and Microsoft's My Phone have been adopted in past collections(Sotillo, 2010;Walkowska, 2009).

(iii) Central collection point: contributors forward messages to a collection number which can be researcher's own mobile phone. To encourage message donations, donors are compensated for their contribution to offset any incurred cost, thus the large-scale collection can be expensive. Messages can also be forwarded to a designated toll free mobile phone operators number in collaboration with researcher in order to motivate contributors (Durscheid and Stark, 2011).

For collection of wide range of SMScorpora, a combination of one or more of these techniques is usually adopted and as technology advances faster and cheaper techniques are expected to emerge.

III. RELATED WORK

The widespread usage of SMShas elicited an avalanche of active research areas such as in studies like sociolinguistic,diachronic and linguistic studies. SMS data plays a vital role in normalization of text. For content based SMS spam filters, SMS dataset is a key tool in training of machine learning algorithm. The demand for SMS corpora is of great benefit.A sizable number of effortshave been made in the collection of SMS corpora.Worth mentioning are the following:

A total of 10,117 messages were collected from 166 university students of the National University of Singapore (How and Kan, 2005). Simple transcription of messages was adopted as collection method and the result of study was made public. Aim of the study was on improving predicted text entry.

The sms4Science project collected 30,000 French text messages from 3,200 participants within the age bracket of 12 – 65. The program was broadcasted on Belgium national television and participants were requested to send their SMS to toll free researchers numbers. The project was aimed atcollecting and transforming the corpus in order to be used as a reference corpus, and translating the language into standard French for future study(Fairon and Paumier,2006).

A publicly-available large-scale multilingual collection of two-sided, naturally-occurring SMS and chat data was done(Song et al, 2012). It was made of over 6.5 million words of everyday chat and SMS in Chinese, Egyptian Arabic and English. Two collection methods were adopted. One was based on real-time capture of SMS or chat messages between pairs of consented users and the other was voluntary donations of archived SMS or chat messages. The dataset was for the evaluation of machine translation systems.

A public live SMS corpus was created (Chen andKan,2012). It comprised of 41,790 English messages and 30,020 Chinese messages. Amazon's Mechanical Turk platform, (Mturk) crowd sourcing strategy was employed to recruit contributors from a wide range of sources .They also adopted web based transcription, SMS export and SMS upload techniques to broaden the breath of the corpus.

A SMS spam corpus was created by collecting 450 SMS legitimate messages from Caroline Tag PhD Thesis, 3375 SMS legitimate messages from NUS SMS corpus, 425 SMS spam messages from Grumbletext Website and finally 1,002 ham and 322 messages from V.0.1 Big SMS spam corpus(Almeida Gómez Hidalgo and Silva, 2013). A collection of 4,827 Ham messages and 747 spam messages totaling5,574 messages was collected. A near-duplicate detection method was employed to detect duplicated messages. The corpus was used for the validation and classification of SMS spam filters.

Working on the framework of the SoNarDutchproject, 53,000 Dutch text messages was collected from 272 participants. A website was setup for contributors to donate SMS through the internet and software that could be used to upload SMS from donors phones were made available for the Android, apple iphone and Nokia. The Dataset was for sociolinguistic and normalization studies.

A Dutch Corpus of Facebook Posts and WhatsAppChats was created (Verheijen and Stoop, 2016).Participants were Dutch youth between the ages of 12 and 23 years. A total of 94 Facebook and 34 chats participants contributed 171,693 words to the corpus. Two website was created for the collection of messages,one for chats and the other for Facebook contributors. Sociolinguistic information was the research objective.

IV. ETHICAL CONSIDERATION IN SMS COLLECTION

The personal and private nature of SMShas created a barrier in the collection and creation of SMS corpora. Contributors and donors of text messages are always skeptical and unwilling to part with their SMS messages and when they eventually do, researchers do not get their full cooperation. In order to gain the confidence and cooperation of participant, assurance of the protection of rights and interests of the message donors and any other persons mentioned in the text messages should be guaranteed.

The protection of rights and interest of participant forms the basis of ethical consideration in SMS collection. In ensuring compliance of protection of rights and interestof participants in SMS corpora creation, the following ethical procedures backed by existing legal and ethical guidelines from British Association for AppliedLinguistics (BAAL), The Association of Internet Research (AoIR) and UK Data Protection Act (1998)were considered: participant consent, data anonymization, and ensuring participants' safe information storage.

A. Participant Consent

Participants in SMS collection projects have the right to give informed consent (Oates, 2009).Section 2.1- 2.3 of the BAAL (2006) guidelines clearly spelt out the responsibility of researcher to participants. (AoIR) suggest that participants should be approached at the beginning of a research projects and briefed on their rights to confidentiality, privacy as well as informed consent if their right to privacy is to be protected (Rock 2002:6).

Majority of SMS corpus considered in this review obtained consent of participant in one way or the other before commencement of collection or donation of text messages.

In Tagg (2009) thesis, two stages of consent was obtained; firstly, through verbal request of participants initial informed consent and secondly, through written consent by all message donors that participated.

All contributors in Chen &Kan (2012) public live SMS corpus were pre-informed about the intention to publish the resultant corpus and make it publicly available online before the commencement of the project.

A customized page within LDC's WebAnn framework for users to sign up and provide their consent to participate and enroll in the collection was provided (Song et al., 2012). The consent form assured participant of the following:

- Participants will not be recognized by name, phone number, email address, chat ID in completed corpus or any other corpora, presentations or publications related to the project.
- All personal information required for the project will be stored in a private and non-sharable database.
- Message headers that contain phone numbers, chatIDs, email addresses or other personal information are deleted before the message content is added to the corpus.

- Dissatisfied participant can withdraw from the exercise and can equally request for removal of donated messages and conversation at any time.

To assure contributors of their privacy and interestTreurniet, DeClercq, Heuvel and Oostdijk(2015) in their Dutch SMS corpus, sent the following instructions to contributors before they uploaded their messages "To protect your privacy, we are removing sensitive information in your SMS. This process is done on your device, so your SMS is not sent to our server yet. Despite this process, you may want to have a look at the messages below and remove messages you do not wish to donate. To do this, just remove the text between the dividing lines (----)." Phone numbers and other private data inside their messages were replaced with a unique identifier.

To safe guard the authors right and interest, Verheijen and Stoop(2016) obtained by consent the Intellectual property rights (IPR) of both Facebook and individual contributors of Facebook and WhatsApp messages. Parents or guardian of contributors between the age of 12- 17 were made to sign the Web consent form before the acceptance of their messages.

B. Anonymizing Data

SMS comprises both personal and intimate information, such as account numbers, email addresses and phone numbers. "Therefore message donors have the right to remain anonymous. Their confidentiality should be respected, and an attempt made to anticipate potential threats to both anonymity and confidentiality (e.g. by anonymizing the data, making it secure, and sometimes even destroying it)"as stated in section 2.4 of the BAAL(2006) guide;ACFID(2016).To ensure the maintenance of privacy, anonymization (substitution of all names and contact details relating to specific individuals who could subsequently be identified.) as promised in the consentform must be implemented. Methods for anonymizing data can be manual, automatic or a combination of both (Rock, 2001).
The public live SMS corpus adopted DES encryption to create a one-way enciphering of the phone numbers, which replaced the originals in the corpus while the body of the message that contain sensitive data such as dates, times, decimal amounts, and numbers with more than one digit (telephone numbers, bank accounts, street numbers, etc.), email addresses, URLs, and IP addresses were replaced automatically with semantics placeholders(Chan and Kane, 2012). Table 1 depicts replacement codes used for anonymization process.

TABLE1. REPLACEMENT CODES USED FOR ANONYMIZATION PROCESS.

Original Content	Example	Replaced Code
Email Address	name@gmail.com	$\langle EMAIL \rangle$
URL	http://www.google.com	$\langle URL \rangle$
IP Address	127.0.0.1	$\langle IP \rangle$
Time	12:30	$\langle TIME \rangle$
Date	19/01/2011	$\langle DATE \rangle$
Decimal	21.3	$\langle DECIMAL \rangle$
Integer over 1 Digit Long	4000	$\langle \# \rangle$
Hyphen-Delimited Number	12-4234-212	$\langle \# \rangle$
Alphanumeric Number	U2003322X	$U\langle \# \rangle$X

The Dutch SMS corpus adopted same automatic anonymization process in replacing sensitive data, such as dates, times, decimal amounts, and numbers bank accounts, street numbers and telephone numbers with semantic place holders(Treurniet, De Clercq, Heuvel and Oostdijk, 2015).

In an effort to protect and conceal participants personal information, (Tagg, 2009)used a semi-automatic anonymization process toreplacepersonal names and contact details of individuals with codes and numbers. The process involved thegeneration of case-sensitive word-frequency list from which words beginning with upper case letters were extracted and secondly, the removal of duplicates of the same nameform.

The BOLT Phase 2 Corpus also ensured that participants were never identified by name, phone number, email address, chat ID or other personal identifier in the corpus, publications or presentations(Song et al.,2012). This was achieved by the automatic deletion of message headers that contain phone numbers, chatIDs, email addresses or other personal information during message processing stage.

C. Safe Data Storage

Corpora must be stored in such a manner that guarantees participants 'rights are protected in accordance with the UK Data Protection Act (1998);ACFID(2016) and that participants are informed of this(BAAL, 2006). Appropriate measures should be taken to store participant data in a secure manner and to achieve this, consent forms should contain such information as description of intended uses, disposal or storage procedures and documentation procedures for data including an option to agree or disagree with these procedures (Tagg,2009).

In Tagg (2009) research, assurance was given to participants in the consent form that "all personal data used in the course of the research shall be processed in accordance with the Data Protection Act 1998"and subsequently shall be destroyed within a period of three years after the research work. The consent form also stated that all messages accepted for

publication would be anonymised before such information are released to the public domain.

Two measures taken in handling safe storage of participant information by (Song et al, 2012) were: Storage of participants' personal enrollment information in a secured non-sharable database and the addition of another layer of protection of participant information at the data upload stage. This enabled participant to edit and choose the final content of what goes to the corpus. To further allay the fears of participant in the safety and protection of their information, Treurniet, De Clercq, Heuvel and Oostdijk (2015) sent the following instruction to participant "To protect your privacy, we are removing sensitive information in your SMS. This process is done on your device, so your SMS is not sent to our server yet" while participants were pre informed of intention of the researcher to store collected messages in a database which will be made available for scientific research according to Radboud University's ethical rules and also ensured that provided data were anonymised in such a way that they are not traceable to the original authors (Verheijen and Stoop,2016).

V. CONCLUSION

As the surge in the usage of SMS and other social media Computer-mediated communication (CMC) is attracting research in linguistic, diachronic, normalization and sociolinguistic studies and also in the validation and comparison of different classifiers in SMS spam filters, the need for the availability of Corpora to enhance these studies is of high importance. This research focused on the different techniques for the effective collection SMS data for the creation of different corpora. It alsoelucidates ethical conditions that motivates participant to part with their private and confidential messages.These shouldact as a guide to future corpus developers.

REFERENCES

[1] Almeida, T., Gómez Hidalgo, J.M., Pasqualini Silva, T. Towards SMS Spam Filtering: Results under a New Dataset.International Journal of Information Security Science, Vol 2, No 1, 2013

[2] Australian Council for International Development (ACFID).Principles and Guidelines for ethical research and evaluation in development.14 Napier Close, Deakin ACT 2600 Private Bag 3, Deakin ACT 2600, Australia, 2016

[3] BAAL (2006).Recommendations on Good Practice in Applied Linguistics'. Retrieved from: http://www.baal.org.uk/goodprac.htm.

[4] Chen, T. &Kan, M. Y (2012). Creating a Live, Public Short Message Service Corpus

[5] Cloudmark whitepaper. SMS Spam Overview. Preserving the value of SMS texting. Retrieved from:https://www.cloudmark.com/en/s/resources/whitepapers/sms-spam-overview

[6] Durscheid, C and E. Stark. ¨ SMS4science: An International Corpus-Based Texting Project and the Specific Challenges for Multilingual Switzerland, Chapter 5. Oxford University Press, 2011

[7] Elizondo, J. Not 2 Cryptic 2 DCode: Paralinguistic Restitution, Deletion, and Nonstandard Orthography in Text Messages. Ph. D. thesis, Swarthmore College, 2011

[8] Fairon, C. and Paumier, S. A translated corpus of 30,000 French SMS. In Proceedings of *Language Resources and Evaluation Conference*.2006,Genova.

[9] GOV.UK. Data Protection Act. Retrieved from: https://www.gov.uk/data-protection/the-data-protection-act.

[10] How, Y. and M. Kan. Optimizing predictive text entry for short message service on mobile phones. In Proceedings of *Human-Computer Interaction Institute(*HCII). Lawrence Erlbaum Associates, 2005

[11] Oates. B.J.Researching Information Systems and Computing.SAGE Publications Ltd, London, 2009

[12] Sanders, E. (2012). Collecting and Analyzing Chats and Tweets in SoNaR. In Proceedings of *Language Resources and Evaluation Conference* 2012,Istanbul, Turkey.

[13] Song.Z, Strassel. S, Lee. H, Walker. K, Wright.J, Garland.J, Fore.D, Gainor.B, Cabe.P, Thomas.T, Callahan.B, Sawyer.A. Collecting Natural SMS and Chat Conversations in Multiple Languages:The BOLT Phase 2 Corpus. Linguistic Data Consortium, University of Pennsylvania, 2012.

[14] Sotillo, S. SMS Texting Practices and Communicative Intention. Hershey: IGI Global, Chapter 16, pp.252–265, 2010

[15] Tagg, C., (2009). A corpus linguistics study of SMS text messaging. Ph.D. thesis, University of Birmingham, united Kingdom.

[16] Treurniet, M., De Clercq, O., Oostdijk, N., Heuvel, H. vanden, (2012) Collecting a Corpus of Dutch SMS. In Proceedings of LREC 2012, Istanbul, Turkey,

[17] Verheijen L., Stoop W. Collecting Facebook Posts and WhatsApp Chats. In: Sojka P., Horák A., Kopeček I., Pala K. (eds) Text, Speech, and Dialogue. TSD 2016. Lecture Notes in Computer Science, vol 9924. Springer, ChamRock, F. (2001) _Policy and practice in the anonymisation of linguistic data' InternationalJournal of Corpus Linguistics 6/1: 1-26.

[18] Walkowska, J. Gathering and Analysis of a Corpus of Polish SMS Dialogues. Challenging Problems of Science. Computer Science. Recent Advances in Intelligent Information Systems, 145–157, 2009.