A MACHINE LEARNING BASED APPROACH TO COMPRESS ENGLISH POSTS FROM SOCIAL MEDIA TEXTS

Ankita Datta(14UCS003)
Manaswita Datta(14UCS031)
Puja Das(14UCS158)
Subhankar Das(14UCS155)

COMPUTER SCIENCE & ENGINEERING DEPARTMENT
NATIONAL INSTITUTE OF TECHNOLOGY,AGARTALA
INDIA-799046
APRIL, 2018

A MACHINE LEARNING BASED APPROACH TO COMPRESS ENGLISH POSTS FROM SOCIAL MEDIA TEXTS

Report submitted to
National Institute of Technology, Agartala
for the award of the degree
of
Bachelor of Technology

by

Ankita Datta(14UCS003)
Manaswita Datta(14UCS031)
Puja Das(14UCS158)
Subhankar Das(14UCS155)

Under the Guidance of

Mr. Dwijen Rudrapal
Assistant Professor, CSE Department, NIT Agartala, India

COMPUTER SCIENCE & ENGINEERING DEPARTMENT
NATIONAL INSTITUTE OF TECHNOLOGY AGARTALA
APRIL, 2018

Dedicated To

To our Project Supervisor Mr. Dwijen Rudrapal, Assistant Professor, CSED, NIT Agartala for sharing his valuable knowledge, encouragement & showing confidence on us all the time. This project would not have been possible to come about without the support and guidance that we received from him. To each of the faculties of the department who contributed in our development as a professional and help us to achieve this goal.

To our classmates who have somehow contributed to the creation of this project and who have supported us.

"Information is the oil of the 21st century, and analytics is the combustion engine."

- Peter Sondergaard (Vice President, Gartnern)

REPORT APPROVAL FOR B.TECH

This report entitled "*A Machine Learning based approach to compress English posts from Social Media texts*", by Ankita Datta (14UCS003), Manaswita Datta (14UCS031), Puja Das (14UCS158), Subhankar Das(14UCS155), is approved for the award of ***Bachelor of Technology*** in ***Computer Science & Engineering***.

Mr. Dwijen Rudrapal	Dr. Rup Narayan Ray
(Project Supervisor)	(Head of the Department)
Assistant Professor	Assistant Professor
Computer Science and Engineering Department	Computer Science and Engineering Department
NIT, Agartala	NIT, Agartala

Date:_____

Place:NIT, Agartala

DECLARATION

We declare that the work presented in this report proposal titled "*A Machine Learning based approach to compress English posts from Social Media texts*", submitted to the Computer Science and Engineering Department, National Institute of Technology, Agartala, for the award of the ***Bachelor of Technology*** degree in ***Computer Science & Engineering***, represents our ideas in our own words and where others' ideas or words have been included, we have adequately cited and referenced the original sources. We also declare that we have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in our submission. We understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

APRIL, 2018
NIT, Agartala

Ankita Datta	Manaswita Datta
Puja Das	Subhankar Das

CERTIFICATE

It is certified that the work contained in the report titled "*A Machine Learning based approach to compress English posts from Social Media texts*", by Ankita Datta(14UCS003),Manaswita Datta(14UCS031),Puja Das(14UCS158) ,Subhankar Das(14UCS155), has been carried out under my supervision and this work has not been submitted elsewhere for a degree.

Mr. Dwijen Rudrapal	Dr. Rup Narayan Ray
(Project Supervisor)	(Head of the Department)
Assistant Professor	Assistant Professor
Computer Science and Engineering Department	Computer Science and Engineering Department
NIT, Agartala	NIT, Agartala

Acknowledgement

We would like to take this opportunity to express our deep sense of gratitude to all who helped us directly or indirectly during this thesis work.
Firstly, we would like to thank our supervisor, **Mr. Dwijen Rudrapal**, for being a great mentor and the best adviser we could ever have. His advise, encouragement and critics are source of innovative ideas, inspiration and causes behind the successful completion of this report. The confidence shown on us by him was the biggest source of inspiration for us. It has been a privilege working with him from last 1 year.
We are highly obliged to all the faculty members of Computer Science and Engineering Department for their support and encouragement. We also thank **Prof. H. K. Sharma**, Director, NIT, Agartala and **Dr. Rup Narayan Ray**, Head of the Department, Computer Science and Engineering Department, NIT, Agartala for providing excellent computing and other facilities without which this work would not achieve its quality goal.

- Ankita Datta

- Manaswita Datta

- Puja Das

- Subhankar Das

List of Figures

List of Tables

Abstract

During recent years, socially generated content has become prevalent on the World Wide Web. The enormous amount of content generated in social networking sites and micro-blogs such as Facebook and Twitter has empowered ordinary users of the Web. This increasing volume of text generated necessitated the need for automatic summarization of social media texts like Facebook posts. This paved the work for sentence compression of such non-standard texts. Such sentence compression can be formally defined as the creation of an abbreviated sentence in a way that preserves the original meaning of the sentence and retains the grammaticality. In this project, we propose an approach for compressing English sentences collected from Facebook by training the model on a parallel corpus to decide which word to delete or retain. We apply four machine learning algorithm (Sequential Minimal Optimization, Naive Bayes, Logistic Regression and Random Forest) to evaluate the accuracy of the compressed sentences over a range of features.

Contents

CHAPTER 1

Introduction

The volume of data on the social media is huge and even keeps increasing. The need for efficient processing of this extensive information resulted in increasing research interest. In this project, we are mainly concentrating on Facebook as a social media texts. According to Statista, the world's largest statistics portal, Facebook is the most popular and fastest growing social network site[1]. There are over 2 billion active users as of January 2018[1]. With nearly 510,000 comments and 293,000 statuses updates per minute[2] , Facebook has become an important source for gathering information on almost every topic. As for example, during FIFA World Cup, 2014, 3 billion posts were generated from June 12, 2014 to July 13, 2014[3].

The posts generated by the user usually contain useful and essential information reflecting the user's opinions. This information could be useful in decision making as well as improving the services or products. These are beneficial for manufacturers or service providers. Moreover, most of people have accepted to read online comments as one of the steps before making a purchase[1]. So, these information could be used for achieving user satisfaction.

[1]https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/
[2]https://zephoria.com/top-15-valuable-facebook-statistics/
[3]https://www.statista.com/statistics/477371/facebook-sporting-events-interactions/

The number of these posts and comments is increasing day by day. Although this rapid growth has many benefits and provides more information, it raises some challenges. The main challenge is accessing the useful and required information in the shortest time. Even interpreting all the posts and comments is quite challenging. There is no way the most significant information can be comprehended.

Since, the systematic analysis of this information is an essential need in so many domains including e-commerce, production, and social network analysis, an automatic summarization system is required that can generate a condensed content. Although various approaches are made for summarization of social media texts, those methods extract good sentences to compose the summaries[2][3]. Those approaches don't shorten the sentence which leads to redundant words and phrases. Our work is motivated to solve this shortening problem.

Sentence compression is a subtask for text summarization. Although, besides automatic summarization of text, sentence compression can be useful for various other Natural Language Processing (NLP) processes like semantic role labeling[4], question generation[5]. The state-of-the-art sentence compression methods are only for standardized texts. Standardized texts are those found in newspaper particle, documents. But the social media texts differ from the former.

The use of social networks has made everybody a potential author, so the language is now closer to the user than to any prescribed norms. Posts, tweets and status updates are written in an informal, conversational tone than the carefully edited work that might be expected in traditional or formal texts. This informal nature of social media texts presents new challenges.

Inconsistent (or absent) punctuation and capitalization can make detection of sentence boundaries quite difficult. Emoticons, incorrect or non-standard spelling, and non-standard abbreviations complicate tokenization and part-of-speech tagging, among other tasks. Traditional tools can't adapt to new variations such as letter repetition ("heyyyyyy"), which are different from common spelling errors. Grammaticality, or frequent lack of one, is another concern for analyses of social media texts. Also, the choice between "there", "they are", "they're" and "their" can seem to be made at random. Moreover, social media texts are commonplace for typological errors and chat slangs.

Contrary to the informal writing that takes a personal approach, and often resembles spoken language, formal writing forms a whole separate category. The language is much more professional and business-like, and is considered as a standard writing style.

Here are a few rules regarding formal writing:-

- Absolutely no contractions. (e.g. can't, don't, won't, etc.)
- Standard punctuation and spelling is expected.
- Must be organized into paragraphs filled with long, complex sentences.
- Writing style must not be casual.

Few comparisons of formal and informal sentence are:-

- Informal: Nice to see you! come again!

  Formal: It has been very nice to see you. We would be glad to see you again soon.
- Informal: Hot as hell! You could fry an egg on the sidewalk.

  Formal: It is extremely hot today.
- Informal:She's decided to accept the job.

  Formal: She has decided to accept the job.

In the first two examples, items which we would normally expect to use in a sentence if we follow the grammatical rules, are left out (in other words, when we don't use). This is called ellipsis. As for example :

I am absolutely sure [that] I have met her somewhere before.

However, in the , third sentence, the words are contracted. Usually a pronoun or noun and a verb are contracted and not are combined, in a shorter form. As for example:

She's[She has] decided to accept the job.

It can be concluded that contractions and ellipsis are common in informal language. Henceforth, it contributes to the failure of the existing state-of-art methods for social media text.

According to Knight & Marcu[6], given an input source sentence of words, a compression is formed by dropping any subset of these words. Good compressions are those which: use fewer words than the source sentence and retain the most important information from the source sentence and remain grammatical.

In this project, we propose an approach to obtain the compressed version of a sentence by dropping of unimportant words. Initially, the posts are split into individual sentences. Each sentence is then tokenized. The tokens are tagged by a semi-automatic technique. Along with that, a parallel corpus containing the compressed form of the source sentence is built for training. The model experiments on four machine learning algorithms. To the best of our knowledge, this is the first work on sentence compression of social media text.

1.1 Motivation

The prime motivation comes from the need to compress informal sentences occurring in social media. During any important event posts are generated at the rate of hundreds to thousands per second. While traditional media provide unidirectional communication from business to consumer, social media sites have allowed interactions among users across various platforms and therefore become a primary source of information for open intelligence.

1.2 Goal

The primary goal is to create a approach to reduce the length of sentences by removing less important words from the sentence collected from Facebook. While attempting to generate output sentences, we are trying to capture the most important elements of the original sentences so that the generated sentence makes sense. Our approach uses a corpus of English sentences collected from Facebook which are manually compressed for training the data set. We use Ritter's POS tagger, and the Weka machine learning package for evaluating the accuracy. Four machine learning algorithm namely Sequential Minimal Optimization, Naive Bayes, Logistic Regression and Random Forest is applied to the dataset to evaluate the accuracy.

CHAPTER 2

Related Work

Several approaches both machine learning based and rule-based have been proposed till date. But all of them experiment on standardized or formal texts. Our work is conducted over social media texts or more precisely, Facebook posts. However some of the works are mentioned below.

2.1 Sentence compression for formal text

Some researchers[7][8] used sentence compression to reduce redundancy in generating summaries while others used it to generate TV subtitles or text for small screens[9][10].

Some models formulated the task as a word or phrase deletion problem. For e.g., Knight & Marcu[6] proposed two models for sentence compression: a probabilistic noisy channel model and a decision-based model to generate abstractive summaries by dropping a subset of words. The probabilistic model compressed the sentences using Naive Bayes rules.The deletion problem was further improvised to address global dependency in deleting local words[9][10].

A few sophisticated models were introduced to improve the accuracy further. Two sentence reduction algorithms were proposed by Nguyen & Ho[11] – a template-based model and a Hiddden Markov Model. The template-based learning model learns lexical translation rules and a Hidden Markov Model learns which sequence of lexical rules should be applied to a sentence. The model performed better than Knight & Marcu [6] based on grammaticality and importance measure.

Filippova & Strube[9] used rules for syntactic and semantic modification of parse tree to generate the compressed form. Cohn & Lapata[12] gave a compression method that extracts tree transduction rules from aligned, parsed texts. Then using a max-margin learning algorithm learns weights on transformations. Later, they[13] presented tree-to-tree transducer that is capable of transforming an input parse tree to a compressed parse tree based on Synchronous Tree Substitution Grammar. Lexicalized Markov Grammar was used by Galley et al.[14] for sentence compression. Clarke & Lapata[15] used Integer Programming approach and encoded various decision variables and constraints in the model to guarantee the grammaticality of the reduced sentence.

The unsupervised model proposed by Cordeiro et al.[16] used Inductive Logic programming.(ILP) Banerjee et al.[17] used ILP for the generation of abstractive summaries by sentence compression. Most recently, Lu & Liu [18] presented an approach based on a Re-read Mechanism and Bayesian combination model for sentences compression.

2.2 Sentence Compression for Informal Text

To the best of our knowledge, this project is the first attempt for compression of informal or social media texttext i.e. English sentences collected from Facebook.

CHAPTER 3

Importance and challenges of Facebook texts

Social networking websites play a vital role in user's online activities. The Worldwide Internet users have increased rapidly. Day by day the amount of user generated content has exponentially increased on these social networking platforms. These social media contain text written in single, multiple languages or even code-mixed. When people write on social media websites they don't look up on the sentence formation, spelling, grammatical errors etc. Because of these, many linguistic ambiguities occur in social media platform.

3.1 Importance of Facebook posts

Facebook is by far the largest social networking site. The two thirds of Facebook users get news there[4]. Study of news consumption on Facebook found users are experiencing a relatively diverse array of news stories on the site.

Some of the most important roles of Facebook posts is mentioned below:

[4]http://www.journalism.org/2016/05/26/news-use-across-social-media-platforms-2016/

- Crisis-related situation: Social Media like Facebook has become a valuable means of communication in many places affected by a natural disaster, which allows people to keep in touch with family and friends and access important information.

- Politics: Social media like Facebook is stronger and more persuasive than television in influencing people. The political leaders have millions of supporters on Facebook. It plays a huge role and influence in elections.

- Sports: Facebook is also playing an active role in the sports like Indian Premier League (IPL) by providing up to date and live information. Teams are in constant touch with their fans through it and there is great interaction. For IPL sponsors and brands, such an interaction and reach of social media is a boon.

- Business[5]: For business purposes, Facebook is the most important social media platform as there is customer's engagement. It plays an important role for marketing because of its popularity, content capabilities, and advertisement targeting tools. Facebook collects details about its members than other services which allow businesses to create advertisements by profession, interests and social connections.

3.2 Types of Informal texts

The text written in news paper, letters, articles or anywhere for formal communication are called Formal Text, while texts used on social media are called Social Media Text or informal text . Social media texts can categories into two categories: monolingual texts and Bilingual or Multilingual texts.

3.2.1 Texts in Monolingual

Monolingual texts are written only in single language using Unicode or Phonetic Typing. One monolingual text is mentioned below:

I am absolutely sure that I have met her somewhere before.

[5]http://smallbusiness.chron.com/importance-facebook-marketing-38546.html

3.2.2 Texts in bilingual/multilingual

Multilingual texts or code-mixing is the mixing of two or more languages. Multilingual speakers increase the chance to mix two or more languages which arises writing in bilingual or code mixing. A code-mixed text is mentioned below:

Sarkar keh rahi hai paisa ki koi kami nahi hai, but still there is no cash in ATMs.

3.3 Challenges of Facebook texts

Variation in writing style of different people causes a number of challenges which are discussed below.

- Non Structured Data : The contents available on facebook are in non structured form. The different available formats are text, video, audio, image etc.
- Non Standard Abbreviation : A new trend is increasing, that the users write comments in non standard abbreviations like ‘gm’ for ‘good morning’. The variation in abbreviation of words or phrase depends on person.
- Typological Errors : Anyone can frequently and correctly writes using pen and paper. But when they use keyboard to write anything then chances of typing error increases, which is a very common issue.
- Repetition of Characters : When writers want to emphases on a word then they simply repeats the characters within the words . In word “okkkkkk”the character ”k“ is repeated four times to make emphases on original word “ok”.
- Cognitive Error[10] : Cognitive errors occur when user writes wrong word in place of correct word according to that sentence in which word is used. The spelling of word is correct but only the meaning or sense of word is wrong according to the sentence. For example:

  I like **there** living style.

  Here, word “there” is wrong word according to the sentence, the correct word is “their”.

- Multiword Tokens[10]: Users write a single word in place of multiple words like ‘asap’ for multiple words “as soon as possible”. This has become new writing trend on social media to use short cut.
- Creative Use of Punctuation : Some people on social media creatively use Punctuation to express their feeling. For example, :-) for happiness and :- (for sadness.
- Non-Linguistic Sounds [10]: The internet users express their feelings (happiness or sadness) using lexical term, which are not the part of any language. For example, “hahaha” for laughing.

CHAPTER 4

Corpus formation And Annotation

In this section, we describe the methodology adopted for collecting posts, selecting the required posts, tokenization of the corpus, preparing the parallel corpus and tagging the tokens.

4.1 Collecting posts

The first step was to collect posts from Facebook. For this purpose, we used Facebook API[6] to fetch a total of 1884 posts related to UN election 2017 (CNN Facebook page[7]).

After a manual inspection, few posts were removed, being code-mixed or being non-English, as we restrict to English posts only. From these remaining posts, the first sentence of each was extracted since our task is related to sentence compression. Soon it was discovered that some sentences were identical. Those sentences were removed to make the posts non-redundant. Some sentences were also dropped because of the following reasons:

[6]https://developers.facebook.com/
[7]https://www.facebook.com/cnn/

- Some sentences didn't convey any meaning. For example: "@tmkd had a laugh today with plq audience for @flisee tune intended reaction?"
- Two sentences seemed like a single sentence because of recursive periods or lack of end markers. For example, "We are the majority and we demand dem.....help spread this message far and wide******"

Finally, 1200 consistent and meaning generating sentences were manually selected for further processing.

4.2 Annotation Process

The collected data were annotated using a semi-automatic technique to speed up the manual tagging. Tokenization is difficult for social media texts because of the noisy and informal nature. Ritter's tagger[8] was employed to tokenize and tag the data in the first place. Although developed for tweets, it works even for Facebook posts since the nature of both the texts are similar.

While analyzing the corpus, it was found many sentences were not tokenized properly, as expected. This was mainly because of lack of white space between words (e.g. 1). Sometimes a word was tokenized into two because of use of punctuations (e.g. 2). All these complications were resolved manually.

1. In the sentence, *"Because he legitimately hasn't passed 1piece of legislation....", '1piece'* was treated as a single entity and was separated into two entities:*'1', 'piece'*.

2. In the sentence, *"Leftists should like N Korea, Gov't runs everything", 'Gov't'* was divided into two tokens: *'Gov' and "t'*. This was combined to form a single token.

From this 1200 sentences, a total of 16445 tokens were formed i.e. approximately 13.70 tokens for each sentence. The tagger also detailed the POS tag, the chunk and the name entity for each token.

However, the tags generated by Ritter's tagger lead to some erroneous tagging. To overcome those erroneous tagging manually inspection of the whole corpus was done by two annotators. For this purpose, the annotators were provided with the same file and instructed to manually

[8]https://github.com/aritter/twitternlp

check the tag, Name entity and chunk and correct them when errors were detected. In case of disagreement, they were asked to discuss until they reach a conclusion.

For preparing a parallel corpus, an annotator was provided with the same file. The instruction provided to the annotator is as follows:

The main aim of this task is sentence compression. You will be provided with a set of English sentences about the US election 2017 collected from Facebook. You are required to read each sentence and compress it to retain the grammaticality while keeping the most important information intact.

To do so, you are allowed to delete words or phrase from the source sentence. You are strictly restricted to insert, substitute or reorder the words in the sentence. You are not allowed to recreate two sentences from a single sentence when sentences are connected through conjunctions. You're also not allowed to delete any sentence from the original file.

Sometimes you may come across sentences that cannot be shortened. In this case, the compressed and source sentence are identical. Hence, keep the sentenced unedited.

Since there is no ideal compressed form of a sentence, all compressions will be considered valid as long as they retain the original meaning and are grammatically correct. Table I shows two examples of sentence compression.

Source 1:	*Hahahah Todd just contradicted his own argument*
Compressed:	*Todd contradicted his own argument.*
Source 2:	*Because he legitimately hasn't passed 1 piece of legislation...*
Compressed:	*he hasn't passed 1 piece of legislation...*

Table 1: Two samples of sentence compression

The obtained parallel corpus is treated as the gold-standard dataset. It should be noted that all the annotators are familiar with the POS tagger through academic learning.

CHAPTER 5

System Description

This section discusses the various features used for training the classifiers, the various classifiers used and the evaluation metrics adopted for the performance measure.

5.1 Features

Feature selection is the most important task of a machine learning approach. Feature selection, also called variable selection or attribute selection, is the automatic selection of attributes in the data (such as columns in tabular data) that are most relevant to the predictive modeling problem we are working on.

The features used in our approach are mentioned and discussed below:

- The current word (W)
- The previous (PW) and previous to previous word (PPW), next word (NW) and next to next word (NNW).

- The chunk (P): While analysis the parallel corpus, it was observed that either the whole chunk was removed or reduced to its corresponding head word to obtain the compressed sentence. The chunk was identified using Ritter's and manually rectified.
- POS tag of W, PW, PPW, NW, NNW, P: POS tagging is used to resolve the ambiguity that exists between words. The POS tag of the lexical features is chosen to resolve this ambiguity.
- Name-entity (NE): Name entities are nouns and hence kept intact for most of the sentences. These tokens are mostly not removed for compression and hence are identified.
- The head word of the chunk (H): The headword in a chunk is that word which is essential to the core meaning of the chunk. In other words, the chunk can be reduced to the head word. For e.g. in case of a noun phrase, the head work is likely to be the subject.

W	POS_W	P	NE	PW	POS_PW	PPW	POS_PPW	NW	POS_NW	NNW	POS_NNW	H	R/D
Are	NNP	Are	O	NIL	NIL	NIL	NIL	you	PRP	really	RB	Are	R
you	PRP	you	O	Are	NNP	NIL	NIL	really	RB	that	IN	you	R
really	RB	really	O	you	PRP	Are	NNP	that	IN	ignorant	NN	really	D
that	IN	that	O	really	RB	you	PRP	ignorant	NN	?	.	that	D
ignorant	NN	ignorant	O	that	IN	really	RB	?	.	NIL	NIL	ignorant	R
?	.	?	O	ignorant	NN	that	IN	NIL	NIL	NIL	NIL	?	R

Figure 5.1: Feature for the sentence "Are you really that ignorant?"

The former two lexical features are inspired from Jurafsky et al.[19] where the relationship existing between the current word and its surrounding words was examined.

5.2 The Classifiers

A classifier is an algorithm that can predict the labels of unseen data. There are many different kinds of classifiers that are suitable for different problems. Choosing the right one is crucial for the performance of the program. The choice mostly depends on the type and size of the dataset. For this project, a classifier that can detect whether a given word should be deleted or retained was needed.

Since we are dealing with sentences extracted from Facebook, we are interested in classifiers that work firmly for sequential tagging of data. The four classifiers used in this project namely Naive Bayes, Logistic Regression, Sequential Minimal Optimization and Random Forest. Some of their details are discussed below:

A Naive Bayes (NB) classifier is a binary classifier. It is based on Bayes theorem:

$$P(A|B) = P(B|A)P(A)/P(B)$$

This theorem shows a way to find the probability of A given evidence B. In classification, this means that we have an object A and evidence B. A certain evaluation method is chosen to determine whether the objects belong to the class or not. This classifier considers all features to be independent.

A Logistic Regression(LR) model is a binary classifier. It is similar to linear regression, where the task is to ascertain a value for each object in the dataset. In contrast to the outcome of linear regression, the outcome of logistic regression is not continuous but binary. A logistic function is used to determine the probability of a value belonging to the class or not.

Sequential Minimal Optimization (SMO) is related to both Support Vector Machines (SVM) and is also popular for text classification.

A Random Forest(RF) classifier averages multiple decision trees based on random samples from the database. A decision tree breaks the dataset down into smaller subsets while simultaneously building a tree with decision nodes and leaf nodes. A leaf node represents a category.

Empirical study shows that NB[20][21] , LR[22], SMO[23] and RF[24] is quite effective for sequential tagging of data. And hence they are chosen.

5.2.1 Evaluation Metrics

Often, accuracy is used to measure how correct the prediction of a classifier is. However since we have an imbalanced dataset, it will not be helpful. We, therefore use F-measure to evaluate the performance of our proposed approach.

CHAPTER 6

Experiment Result

6.1 Result

On the dataset, 10-fold cross validation is performed. In 10-fold cross-validation, the original data-set is randomly partitioned into 10 equal size smaller data-set. Of the 10 smaller data-set, a single data-set is retained as the validation data for testing the model, and the remaining 9 data-set are used as training data. The cross-validation process is then repeated 10 times i.e 10 folds, with each of the 10 data-set used exactly once as the validation data. The 10 results from the folds are then averaged to produce a single estimation. The advantage of this method is that all observations are used for both training and validation, and each observation is used for validation exactly once.

We use Weka[9] for implementing the machine learning algorithms. Weka is an acronym which stands for Waikato Environment for Knowledge Analysis. The Weka machine learning work-bench is a modern platform for applied machine learning. Table 2 describes the evaluation metric obtained after applying the algorithms on the dataset.

[9]https://www.cs.waikato.ac.nz/ml/weka/

It can be clearly observed that SMO gives the highest weighted average F-measure(0.789) while NB(0.770), LR(0.7) and RF(0.759) performed a bit less. It should be noted that SMO and LR is evaluated only on 5000 instances of data-set.

Evaluation Matrics	Clasifier			
	SMO	NB	LR	RF
Weighted average F-measure	0.789	0.770	0.700	0.759

Table 2: F-measure after 10-fold cross validation

When compared with the result previously obtained on 440 sentences i.e Table 3 , it can be seen that for all of the classifiers except LR the F-measure increased. This can be due to the increase in size of the corpus.

Evaluation Matrics	Clasifier		
	SMO	NB	LR
Weighted average F-measure	0.777	0.731	0.752

Table 3: Previously obtained F-measure after 10-fold cross validation

6.2 Discussion

Table 4 lists two compression examples that resulted after compression. Table 5 lists the previously obtained compression sample that was based on a smaller corpus. Contrary to previous results where the grammaticality was not retained, the new results show a substantial increase in retaining the grammar of the sentence. Except for LR, all the classifier resulted in more accurate sentence like the manually compressed sentence. Further analysis showed that for most of the sentences, the length of the compressed version and generated compressed version didn't differ much.

Source 1:	*i personally think that it's disrespectful to the n korean people to go there.*
Compressed:	*it's disrespectful to the n korean to go there.*
SMO:	*it's disrespectful to the n korean people to go there.*
NB:	*that it's disrespectful to the n korean people to go there.*
LR:	*think that it's is to the n korean people to go there.*
RF:	*personally think it's disrespectful to the n korean people to go there.*
Source 2:	*Normally i wouldn't be too happy about the govt telling me where i can and i can't go.*
Compressed:	*i wouldn't be happy about govt telling me where i can and i can't go.*
SMO:	*i wouldn't be too happy about the govt telling me where i can and i can't go.*
NB:	*i wouldn't be too happy about the govt telling me i can and i can't go.*
LR:	*i wouldn't be too happy about the telling i can and i can't go.*
RF:	*Normally wouldn't be too happy about the govt telling where i can and i can't go.*

Table 4: Compression sample

Source 1:	*i personally think that its disrespectful to the n korean people to go there.*
Compressed:	*its disrespectful to the n korean to go there.*
SMO:	*that its disrespectful to the n korean to go there.*
NB:	*its disrespectful to the n korean people to go there.*
LR:	*that its disrespectful to the n korean people to go.*
Source 2:	*Normally i wouldn't be too happy about the govt telling me where i can and i can't go.*
Compressed:	*i wouldn't be happy about govt telling me where i can and i can't go.*
SMO:	*about the govt telling me where i can and i can't go.*
NB:	*wouldn't be happy about the govt i can and i can't go.*
LR:	*i about the telling me where i can and i can't go.*

Table 5: Previously obtained compression sample

However, the proposed approach failed to detect some words accurately. We analyzed those sentences and found the following reasons:

1. In *"Yeah liberals love a nationalist military state and a narcissistic leader with a bad haircut huh"*, lack of end markers and presence of non-linguistic words failed the approach.

2. In sentences with conjunctions, *"CNN is not the only one reporting it but these trolls are paid to troll."*, the conjunction "and" and adverb "not" was not detected properly.

CHAPTER 7

Conclusion & Future Work

We present a machine learning approach to train English sentences from Facebook for obtaining compressed form of the sentences. We detail on collecting, annotating the corpus using a semi-automatic technique and incorporate various features to determine whether a given word should be deleted or retained.

Drawing from the result obtained after applying four different machine learning algorithm (SMO, NB, LR, RF), SMO performed best with an F-measure of 0.789. Being the first attempt at social media text compression, we are able to retain the grammaticality of the compressed sentence to some extent.

Owing to the lack of sentence boundary and use of non-lexical words, proper approach need to be implemented. To tackle the negation problem, global dependency need to be checked. Moreover, the F-measure need to be further improved and generated compressed sentences need to be checked for further efficiency using different other algorithms that are effective for sequential tagging.

References

[1] R. Keshvarian, A. Taei Sadeh, A. Jami, “The role of social networks in online advertising and marketing”. In Proceedings of Second National Conference on Applied Research in Computer Science and Information Technology, 2010.

[2] Evi Yulianti, Sharin Huspi, Mark Sanderson, ”Tweet-biased summarization”. In journal of Association for Information Science and Technology, 2015. 67(6): 1289-1300

[3] Ahmed T. Sadiq, Yossra H. Ali, Mohammad Natiq Fadhil, “Text summarization for social network conversation”. Advanced Computer Science Applications and Technologies (ACSAT), International Conference on IEEE, 2013.

[4] David Vickrey, Daphne Koller, ”Sentence simplification for semantic role labeling”. In Proceedings of Association for Computational Linguistics, 2008: 344–352.

[5] Feras Al Tarouti, Connor McGrory, Jugal Kalita, ”Sentence Simplification for Question Generation”. In Proceedings of International Conference on Computing and Communication Systems, 2015.

[6] Kevin Knight and Daniel Marcu, “Summarization beyond sentence extraction: A probabilistic approach to sentence compression,” Artificial Intelligence, vol. 139, pp. 91-107, July 2002.

[7] Yuya Unno, Takashi Ninomiya, Yusuke Miyao, and Jun'ichi Tsujii, "Trimming CFG parse trees for sentence compression using machine learning approaches," in Proc. of the COLING/ACL on Main conference poster sessions, Sydney, Australia, 2006, pp. 850–857.

[8] Hongyan Jing, "Sentence reduction for automatic text summarization," in Proc. of the 6th conference on Applied Natural Language Processing, Stroudsburg, PA, USA, 2000, pp. 310–315.

[9] Katja Filippova and Michael Strube, "Dependency tree based sentence compression," in Proc. of the 5th International Natural Language Generation Conference, Stroudsburg, PA, USA, 2008, pp. 25-32.

[10] James Clarke., and Mirella Lapata, "Global inference for sentence compression: an Integer Linear programming approach," Journal of Artiicial Intelligence Research, vol. 31, pp. 399–429, Mar. 2008

[11] Minh Le Nguyen, Susumu Horiguchi, Akira Shimazu, and Bao Tu Ho, "Example-based sentence reduction using the hidden markov model," ACM Transactions on Asian Language Information Processing, vol. 3, pp. 146-158, June 2004

[12] Trevor Cohn and Mirella Lapata, "Sentence compression beyond word deletion," in Proc. of the 22nd International Conference on Computational Linguistics - vol. 1, Manchester, UK, 2008, pp. 137–144.

[13] Trevor Cohn and Mirella Lapata, "An abstractive approach to sentence compression," ACM Transactions on Intelligent Systems and Technology, vol. 4, Article 41, June 2013

[14] Kathleen R. McKeown and Michel Galley, "Lexicalized Markov grammars for sentence compression," in Proc. of Human Language Technologies Conference of the North American Chapter of the Association for Computational Linguistics, NY, USA, 2007, pp. 180–187.

[15] James Clarke., and Mirella Lapata, "Global inference for sentence compression: an Integer Linear programming approach," Journal of Artiicial Intelligence Research, vol. 31, pp. 399–429, Mar. 2008

[16] J Cordeiro, G. Dias, P. Brazdil, "Unsupervised induction of sentence compression rules," in Proc. of the Workshop on Language Generation and Summarisation, Singapore , 2009, pp. 15-22.

[17] S. Banerjee, P. Mitra, and K. Sugiyama, “Multi-document abstractive summarization using ILP based multi-sentence compression,” in Proc. of the 24th International Joint Conference on Artificial Intelligence, Buenos Aires, Argentina, 2015, pp. 1208-1214.

[18] Zhonglei Lu, Wenfen Liu, Yanfang Zhou, Xuexian Hu, Binyu Wang, “An effective approach of sentence compression based on re-read mechanism and Bayesian combination model,” in Proc. of Chinese National Conference on Social Media Processing, Beijing, China, 2017, pp. 129-140

[19] Huihsin Tseng, Daniel Jurafsky, and Christopher Manning, “Morphological features help pos tagging of unknown words across language varieties,” in Proc. of the 4th SIGHAN Workshop on Chinese Language Processing, Jeju Island, Korea, 2005.

[20] N. Ramanujam, M. Kaliappan, ”An Automatic Multidocument Text Summarization Approach Based on Naïve Bayesian Classifier Using Timestamp Strategy”. The Scientific World Journal, 2016.

[21] A. Najibullah, ”Indonesian Text Summarization Based on Naive Bayes Method ”. in Proc. of the International Seminar and Conference: The Golden Triangle Interrelations in Religion, Science, Culture, and Economic, Semarang, Indonesia, 2015.

[22] Georgiana Ifrim, Gokhan Bakır, Gerhard Weikum, ”Fast logistic regression for text categorization with variable-length n-grams”. In Proc. of the 14th ACM International Conference on Knowledge Discovery and Data Mining, pages 354–362, 2008.

[23] Majed Ismail Hussien, Fekry Olayah, Minwer AL-dwan, Ahlam Shamsan, ”Arabic text Classification using SMO, Naive Bayesian, J48 algorithms”. In the International Journal of Recent Research and Applied Studies, 9(2), 2011.

[24] B. Xu, X. Guo, Y. Ye, J. Cheng, “An Improved Random Forest Classifier for text categorization”. In the Journal of Computers,7(12):2913–2920, 2012.

CHAPTER 8

Biographical Sketch

Ankita Datta

Surjyamaninagar, Sadar PIN-799022, E-Mail: ankitadatta.agt@gmail.com, Contact. No. +91-8974706832

- Pursuing B.Tech. in Computer Sc. & Engg. branch from N.I.T,Agartala with CPI of 8.13/10/00.
- Intermediate from Netaji Subhas Vidyaniketan,Agartala under (T.B.S.E), Tripura with 76.8% in 2014.
- High School from Netaji Subhas Vidyaniketan,Agartala under (T.B.S.E), Tripura with 86.4% in 2012.

Manaswita Datta

Abhoynagar, Agartala PIN-799005, E-Mail: dattamanaswita@gmail.com, Contact. No. +91-9436996146

- Pursuing B.Tech. in Computer Sc. & Engg. branch from N.I.T,Agartala with CPI of 9.06/10/00.
- Intermediate from Hindi Higher Secondary School ,Agartala under (C.B.S.E), Tripura with 90% in 2014
- High School from Bhavan's Tripura Vidyamandir,Agartala under (C.B.S.E), Tripura with 95% in 1994.

Puja Das

R.K.Mission Road, Agartala PIN-799001, E-Mail: puja250196@gmail.com, Contact. No. +91-8974446727

- Pursuing B.Tech. in Computer Sc. & Engg. branch from N.I.T,Agartala with CPI of 7.44/10/00.
- Intermediate from Netaji Subhash Vidyaniketan,Agartala under (T.B.S.E), Tripura with 71.04% in 2014.
- High School from Netaji Subhash Vidyaniketan,Agartala under (T.B.S.E), Tripura with 81.4 % in 2012.

Subhankar Das

Kakraban, Udaipur, PIN-799105, E-Mail: subhankardasnitacs@gmail.com, Contact. No. +91-9485376695

- Pursuing B.Tech. in Computer Sc. & Engg. branch from N.I.T,Agartala with CPI of 6.28/10/00.
- Intermediate from Jawahar navodoya vidyalaya under (C.B.S.E), Tripura with 81.5% in 2014.
- High School from Jawahar navodoya vidyalaya, under (C.B.S.E), Tripura with 9.6/10/00 in 2012.