Contextually Request Tweets Classification Based on Deep Features

By

Muhammad Danish

00000171720

Supervisor

Dr. Sharifullah Khan

Department of Computing

A thesis submitted in partial fulfillment of the requirements for the degree of MS(IT)

In

School of Electrical Engineering and Computer Science,
National University of Sciences and Technology (NUST),
Islamabad, Pakistan.

(July, 2019)

THESIS ACCEPTANCE CERTIFICATE

Certified that final copy of MS/MPhil thesis written by Mr. Muhammad Danish, (Registration No 00000171720), of MSIT-17 (School/College/Institute) has been vetted by undersigned, found complete in all respects as per NUST Statutes/Regulations, is free of plagiarism, errors and mistakes and is accepted as partial fulfillment for award of MS/M Phil degree. It is further certified that necessary amendments as pointed out by GEC members of the scholar have also been incorporated in the said thesis.

Signature:_____

Name of Supervisor: **Dr. Sharifullah Khan**

Date: _____

Signature(HOD):_____

Date: _____

Signature(Dean/Principal):_____

Date: _____

Approval

It is certified that the contents and form of the thesis entitled **'Contextually Request Tweets Classification Based on Deep Features'** submitted by **Muhammad Danish** have been found satisfactory for the requirement of degree.

Advisor: **Dr. Sharifullah Khan**

Signature: _____

Date:_____

Committee Member 1: **Dr. Muhammad Shehzad Younis**

Signature:_____

Date: _____

Committee Member 2: **Mr. Maajid Maqbool**

Signature:_____

Date: _____

Committee Member 3: **Dr. Madiha Liaqat**

Signature:_____

Date: _____

Dedication

To my parents,

Without whom this success would not be possible.

Certificate of Originality

I hereby declare the submission of this thesis as my own work and to the best of my knowledge it contains no particular substance previously published or written by another author, nor material which to a substantial extent has been accepted for the award of any degree or diploma at SEECS, NUST or at any other educational institute, except where due acknowledgement has been made in the thesis.

I also declare that the content of this thesis is the product of my own thesis, except for the assistance from others in the project's formulation, conception or in style, presentation and linguistics which has been acknowledged.

Author name: **Muhammad Danish**

Signature: _____

Acknowledgement

I am grateful to ALLAH who granted me success and it was impossible without His blessings. I would like to give special thanks to my Supervisor Dr. Sharifullah Khan and GEC members for guiding me throughout this work. I would also like to thank annotators for helping me in this work. I want to thank all my lovely friends and family for supporting throughout the degree. I would like to thank everyone who directly or indirectly contributed in this journey.

Muhammad Danish

Contents

List of Figures

List of Tables

Abstract

Millions of tweets are posted on Twitter related to different events and situations. As a result, many research problems have been identified for Twitter data. One of them is tweet classification, especially classification of tweets posted during critical situations, disasters, or in the time of need. Request classification is a sensitive issue as it helps to save precious lives. In previously proposed methodologies, a lot of feature extraction techniques have been used for tweet classification. These techniques lack a lot of aspects that should be considered for request classification. Features extracted by those techniques do not contain any contextual information of the text. Moreover, features are limited to word level. Another issue is the curse of dimensionality due to n-gram features. Metadata is also used as features in some of the proposed methodologies and it did not contribute to the performance of the framework. Improvement in feature extraction methods can lead to better results. This work aims to develop a deep learning-based framework which extracts contextual features from tweets. These contextual features are more reliable, avoid the curse of dimensionality and capture semantic information. Moreover, sentence-level feature has also been extracted from tweets. Three different feature sets have been extracted to achieve maximum output. Multiple classifiers have been used to validate the performance of this framework. In the end, 612-dimensional hybrid features have been created and experimented using four different classifiers. Neural Networks outperformed other classifiers and produced improved results as compared to the baseline model. The proposed framework achieved 89% accuracy, 96% precision, 90% recall, and 93% F1- score using hybrid features with Hold-out validation.

Chapter 1

Introduction

This chapter started with the motivation and background where social media has been explored. Then problem statement and issues are discussed which are not yet addressed in tweet classification. After problem statement, research objective is discussed that will be accomplished in this work. The proposed solution is explained briefly how a system will be developed for request identification. In the end, the thesis outline is stated.

1.1 Motivation and Background

Social media is becoming a key source for news, updates, reviews, requests and to expand social circle. Twitter and Facebook are the ones competing others in this race [2]. These social media apps are used by people to search friends, family and connect with them. People also share personal information, seek for help, look for suggestions and learn about upcoming activities or on going events. With the availability of the Internet worldwide, the generation of the new era is attracted to social media and it has become a part of their lives. Twitter has become one of the top social media app used by millions of users. It has been ranked among most-visited websites [3] and titled as 'the SMS of the Internet' [4].

Twitter is used for different sort of interactions. People can post their messages limited to 280 characters on the twitter called, "tweet". These tweets are accessible to all other users on Twitter. It is also used as a social awareness

platform [5]. Tweets also contain other sources of information such as, images, videos, and URLs to other websites. The diversity in twitter is one of the major reasons that has attracted users from across the world. Users can also follow specific sports, celebrities, politicians. These followers will receive all updates and news posted by them.

Twitter provided many features that can be used to express userâĂŹs feelings, expressions and reactions to posts. One of the most popular features is the sign '@'. This sign is used to address someone or to refer someone but it is declared as a non-native feature [6]. Another way to mention an incident or to address a topic, sign '' is used called "hash-tag". It is used extensively to refer any event, news or trend. Hashtags can be created by anyone and can be followed by anyone. The popular or most followed hashtag becomes trending and is made visible explicitly in Twitter's interface for new visitors. Twitter also keeps updating its users by sharing them relevant information and also enables users to help their friends and other people who are in need [7].

Due to efficient response and easy to use interface twitter is not only used for social interaction but also source of information sharing, like question-answers, surveys, discussion platform and to help each other [8]. For example in question-answers, users have further different motivations and intentions. Sometimes Twitter is used for getting suggestions and motivation and sometimes it is used to gather factual knowledge and opinions [9]. In the time of emergencies, social media is used as situation awareness and coordination source [10]. It has been observed that social media has been used at large scale in emergencies and disastrous situations ,e.g., the Haiti hurricane, Typhoon Haiyan and Hurricane Sandy [11]. Since Twitter is used globally, it is one of the fastest ways to communicate during emergencies, especially in the early hours of the critical situations [12]. In Hurricane Sandy disaster, around 20 million tweets were posted. Some tweets were about situational awareness and some of them were specified as requests posted by those people who needed help, either to rescued or any other assistance. [13]. In JapanâĂŹs Tsunami situation [14], researcher has analyzed and reported that people tweet regarding critical situations, warnings, request for help. People also shared problems

they were facing and the situations in their surroundings regardless they were victims or not. Moreover, tweeting in disaster is just not limited to western countries but has reached to developing countries too ,e.g., Pakistan Floods 2010 [15]. Due to the popularity of Twitter, a lot of efforts have been made to utilize it as a source of communication and help. For example in Hurrican Sandy and Haiti disaster, a tweet of texting Red Cross for donations was posted by thousands of users. In 2015 during Chennai Rains, ChennaiRainsHelp was used by both affected and safe citizens for help to provide relief to the victims. Another hashtag PorteOuverte in the same year was used in France to provide relief to the people who had their flights canceled due to heavy rainfall [11].

So it shows, how important a social app can play a role in the lives of people. It would be fair to say that nowadays if social media is helping people to know each other, share sentiments and views, it can also save a life if properly observed and monitored. A credible system should be developed which can identify actionable and emergency tweets from twitter. One way is to manually read tweets and check whether these are asking for help or not which seems impossible as millions of tweets are posted on a daily basis and cannot be classified manually. The other way is to develop a smart and artificial intelligence-based system which can mine tweets and identify whether they need any kind of respond or not. A lot of work has been done for this noble cause. As the technology and research move forwards it also opens new ways to improve the system and its performance.

1.2 Problem Statement

Tweets are written in a limited available text. In addition, the grammar, syntactic and structural format is not followed. Informal and trending words are used to post a tweet. Symbols and emoticons have been used with words or in place of some words. Most importantly the context of the tweets needs to be considered. Because of the poor structure of sentences and lack of grammar use, complete sentences should be considered for better performance instead

of words. So while mining and assessing tweets, following issues in tweet classification have been seen which should be addressed.

1.2.1 N-gram Features

The major and important issue in text classification is feature extraction. In early years, proposed solutions were unable to perform better due to the use of n-gram model which is an old technique for feature extraction and have many deficiencies ,e.g, scaling, storage, no contextual information. From the definition of n-gram [16], we can see it just assigns the probabilities to the words that occur in sentences or in data. There are many other aspects to consider for features, which has been missed in this model. Another issue in n-gram is curse of dimensionality, which means n dimensions for words. It means each word will be set as a dimension in the model. As dimensions represent scaling, which means an extension in the feature set. In [17], it has been claimed that in many situations simple scaling will not help to make any significant progress. So, it is important to consider this issue and avoid unnecessary scaling.

1.2.2 Syntactic and Semantic Regularities

Syntactic and semantic regularities are the base of any textual content. Syntactic regularity means the reoccurring of words in the structure of the sentence [18]. For example, consider these two sentences, “Please help him with lifting the bag”,“I am running out of cash will you lend me some money, please?”. In both sentences word “please” is used but either it is used in the start or at the end of the sentence. It shows that the word “please” has a specific order syntactically. According to [19], “Semantics is the study of the meaning of linguistic expressions. The language can be a natural language, such as English or Navajo, or an artificial language, like a computer programming language. Meaning in natural languages is mainly studied by linguists.” So semantic regularities mean different words used to depict a scenario, a scene or a situation. In [17], researchers proposed that by preserving syntac-

tic and semantic regularities, the accuracy of the model can be maximized. This shows that in text classification, semantic and syntactic regularities can impact results significantly.

1.2.3 Sentence Level Features

Due to the importance of tweet classification, feature extraction can not be limited to word level prediction as it also gives importance to uncommon words, which are not needed necessarily. Word level features are unable to extract sentence level features which can store more contextual information and a better understanding of the text. Moreover, in word level features, all the words have the same distance from each other that ignores the importance of the specific words which can improve results in the classification. For example, according to [20], some words are repeatedly used for request but due to the word level features, those words were assigned the same probabilities as other common or uncommon words.

1.2.4 Metadata

Sometimes metadata of tweets is also used as a feature set and it appeared that metadata did not help in the performance of the framework [21]. For instance, in [13], additional features like time, URLs, and location of the user were used. But these features did not give expected results. This shows that the nature of the research should be clear in the start.

1.2.5 Time and Resource efficient

Request classification is a sensitive problem which needs quick and responsive actions. With less time and better resource usage, a developed system should be responsive and with the addition of the dataset, it should not take extra resource consumption as well as more time to execute. Its performance should tolerate the scalability of the data and resources. However, the minor requirement can be compromising.

1.2.6 Preprocessing criteria

In Twitter, users do not follow any formal procedure to post tweets. Tweets are posted in informal way. Different symbols and signs are used according to their supposed meaning and these symbols do not have any official credibility. Another issue with these symbols is not used for the same expression all the time. Different users use the same symbols and signs for different meaning and purposes. Which means it can create ambiguity in tweets from a research point of view. Another problem is the use of repeated words, writing a short form of words, etc. These words do not contain any meaning full information but take extra space and time to process. Different non-ASCII characters i.e. æ and §§§ are also part of tweets. These characters may have special meaning visually but in textual consideration, these characters are useless and extra burden to carry.

Twitter also allows mentioning someone in tweets, hashtags, use of numbers or numerical values and URLs. It can be beneficial in research or it can affect results. For example, if research is about behavioral analysis then mentioning in a tweet can be fruitful as it will have responses but if the aim is to find credible information on Twitter, mentioned names can create issue in the classification of credible data. Likewise, if the topic of the research is to check the connection between Twitter and news channels then URLs can be very important as features. Because URLs contain links to different news channels but if research is about sentiment analysis, there is no need to consider URLs in the feature set. This show preprocessing can impact results and performance significantly depending on what type of research is under process.

1.3 Research Objectives

Techniques used earlier in request classification have deficiencies and limitations which are discussed in critical analysis. Our objectives are to implement techniques which are capable and reliable in tweet classification. To use the contextual information of the data, deep features should be extracted which

contains not only word level information and their semantic relationship but also store sentence level features. To manage the curse of dimensionality, a system should be built which can tackle the scaling problem of the dataset as the size of the data can be increased with time. To avoid metadata features overhead, sentence level features should be extracted. To filter tweets from irrelevant and raw data, proper preprocessing steps should be taken as without preprocessing, performance of the system can be compromised. To get optimized results, suitable classifiers should be identified and experimented which can also manage imbalanced data. To reduce resource consumption and computation time, techniques that are time efficient and have speedy execution process should be implemented. To enhance performance, hyperparametric tuning should be done.

1.4 Proposed Solution

Request identification can help individuals during disasters and critical situations. So for better classification, a dictionary of the dataset using [22] was built. This dictionary not only stores semantic and syntactic information but also groups similar words. It also defines the relationship among those words which will help to identify target class. Use of sentence level features based on deep learning is a major breakthrough in text classification and stores information at the sentence level and does not require any additional features ,e.g., metadata [23]. For contextual information, deep level features are extracted which not only store word-level semantic and syntactic regularities but also store the context of the sentence which in return can boost the system‘s performance and yield better results. Linear Regression and Random Forest classifiers are used for request identification. For fast execution process, XgBoost classifier is applied which can perform classification with minimum time. A classifier, Neural Network is used that learned syntactic and semantic word level features, contextual sentence level features without any extra metadata information. In the end, hyperparametric tuning is applied using Grid Search to optimize classifiers and to obtain best-expected results. For eval-

uation and assessment of results obtained from experiments four evaluation matrics, Precision, Recall, Accuracy, and F1 are used. 10-fold cross-validation techniques are also applied for comparison and verification purposes.

1.5 Thesis Outline

The thesis is arranged as follows: Chapter 2 is literature review and will give a thorough introduction to previous work that has been carried out in recent years. It will also contain an overview of some frameworks considered as a foundation in this research line. In Chapter 3, dataset, its annotation process and experiments with implementation details are discussed. Chapter 4 is about evaluation, validation and results obtained from different experiments. In chapter 5, summary and outcomes of the research are discussed. Limitations and future work is also discussed in the same chapter.

Chapter 2

Related Work

This chapter discusses and evaluates previous research that has been conducted in text classification. The section 2.1 discusses how Twitter is being used more than just social application. Then we look into some previous works including, how research has been carried out through the years and what progress has been made. We will discuss different approaches and methods used for request classification. Discussion about techniques for feature extraction and selections that were used in previous work. Machine learning based classifiers are explained with experiments and limitations in previous research. In the last section, a critical analysis and comparison of the proposed frameworks used in the previous studies has been addressed.

2.1 Twitter in Emergencies and Disasters

In the early years of social media, the purpose was to connect and share personal and social content. Later social media applications became multi-purpose sites. Users started to share news, blogs, updates, incidents and other critical situations. In addition, the use of these sites increased enormously. This showed great potential for researchers and a new era of research started. Some aspects associated with research from the perspective of social media are question identification [24–27], content classification [9, 28, 29], and opinion mining [30].

Research on request identification has been a topic of interest for many years. In [21], researchers worked on a dataset of disaster Hurricane Sandy that occurred in 2012. The purpose was to build a system which can predict or classify tweets into their respective class ,e.g., request or not request. Other normal tweets were collected for the same duration, when disaster occurred. Textual features were extracted using n-gram model with other features ,e.g., location, hashtags, URLs, and metadata. Initially, three classifiers SVM, Decision Tree, and Random Forest were trained. Decision Tree classifier gave better results than the other two. This work was a good effort but it has some limitations. Metadata of tweets were used as additional features which did not help in performance improvement but increased overhead. Textual features were extracted using n-gram model which can cause the curse of dimensionality. Another major drawback was that it did not extract contextual features from the text which could have improved results.

A thorough study has been established in [31] regarding speech acts on Twitter. According to their problem statement speech act is a multi-class problem. A baseline was created by identifying and defining a set of rules and regularities which represents different speech acts in tweets on Twitter. These speech acts were defined as recommendation, assertion, expression, miscellaneous, question, and request. Recommendation act means if someone has recommended a person, a website, a product or suggesting something with personal experience. In assertion, a resourceful claim is given about any event, incident, or place. It can be said by authorities, anchors, or by the public. An expression is an act in which something has been appraised or emotions (anger, happy, sad) are shown. The miscellaneous act is defined as a tweet that is not giving any information or did not fit in any other action. Question is a simple understandable act, it is identified as if any type of question is asked or any type of information is required. In the request act, different words are observed which are used in a polite way ,i.e., would you like to help, please, if you do not mind. In figure 2.1, Six identified speech of acts are shown with examples. Then a large corpus of tweets was annotated. Tweets were annotated by experts under defined rules and regularities. In feature

extraction, the n-gram model is used with speech act verbs. Speech act verbs contain 229 English speech acts which are further divided into 37 groups. Four classifiers logistic regression(LR), support vector machine(SVM), naÃŕve base (NB) and decision tree (DT) are used to test and train their dataset. 20-fold cross-validation is used to evaluate the results. But this work was not scalable due to rule-based approach. Moreover, the semantic and syntactic features were missing. Feature extraction used in this model was limited to the word level.

Speech Act	Example Tweet
Assertion	authorities say that the 2 boston bomb suspects are brothers are legal permanent residents of Chechen origin - @nbcnews
Recommendations	If you follow this man for updates and his opinions on #Ferguson I recommend you unfollow him immediately
Expression	Mila Kunis and Ashton Kutcher are so adorable
Question	Anybody hear if @gehrig38 is well enough to attend tonight? #redsox
Request	rt @craigyh999: 3 days until i run the London marathon in aid of the childrens hopsice @sschospices. please please sponsor me
Miscellaneous	We'll continue to post information from #Ferguson throughout the day on our live-blog

Figure 2.1: Six Speech Acts identified on Twitter with examples

Some real-time systems were proposed, these systems not only identify responsive tweets but also improve their performance as more data is added. In [32], system is divided into two major functionalities, first is to check whether a tweet is informative or not. In second step, classify these informative tweets into two or more classes ,e.g., 'damage' and 'need'. The system consists of three components. In first phase, tweets are collected from Twitter using an API provided by the Twitter Management team. It focuses to collect disaster-related and informative tweets. To collect specific tweets, the system used a set of keywords and defined coordinates where the disaster occurred. After tweet collection, tweets are passed to a group of annotators who label these whether tweets lie in request class or in other class. Finally, these tagged

tweets are passed to classifiers which learn and predict the new tweets. The system used the N-gram model with uni-gram and bi-gram as features to feed it to the classifier. The system supposed to improve its results as it was implemented in real time but there were some issues in the proposed methodology. A major issue with the system was scalability, n-gram is used for feature extraction which scales with the data. In result, system can require more storage and processing resources. The models used to extract contextual features are not used [22, 23].

Question mining is also a hot topic in social media, especially in Twitter as it contains short tweets. Different approaches have been adopted for question identification and classification on Twitter. A pipeline based on multiple tools was developed for question identification by researchers in [26]. To handle the linguistic complexities of the language, parser, tokenizer, and customized lexicon tools were used to identify tweets with questions. To handle repeated words ,e.g., 'pleeeese' or 'whatttttt', the tokenizer was improved which also differentiates various emoticons used in tweets. A parser was used to detect questions using 500 context-free rules. A different set of rules used to classify questions in a rule-based approach. Machine learning approach was also implemented along with the rule-based approaches to identify question tweets. In another framework, a two-stage approach was adopted [25]. It opted better technique in which question identification is divided into two processes. In the first phase, tweets were identified whether they are question or not question regardless of their meaning or purpose. For example "Want 2 kill my boredom! Checked my mobile and the result? Insanity! Low battery" seems like a question but it is not. So to tackle this problem, a second phase was introduced in which all question tweets will be filtered again to check whether these tweets contain meaningful information or not. For example, "What is the name of the author of the Lord of the Rings movie?" is a sure question. After feature extraction, these tweets were trained and tested on Random Forest classifier with 10 fold cross-validation. A limitation was the use of rule-based approach which has become obsolete and even claimed dead by some researchers in 2013 [33]. The hyper parametric tuning was also missing which could have

enhanced the performance of the classifiers.

2.2 Critical Analysis

In existing work, major drawbacks and limitations has been observed and discussed. Another issue with these studies is that the approaches, they have adopted, have become old or replaced or declared dead ,e.g., Rule-Based approach by researchers [33]. This means with the advancement in research, an improved system can be developed with better results and higher accuracy.

Metadata of tweets is also used in previous work as a feature set and some of them accepted that it did not help in the performance of the framework. For instance, in [13], additional features like time, URLs, and retweet were used to identify different classes. These features did not give expected results. The imbalanced dataset can also impact a classifier‘s prediction during testing. If there are multiple classes and one of them has fewer data to train, it may confuse the classifiers and can lead to overfitting on the data set. As we can see in [31], request class results were only 16% because of imbalanced data set. Hence, size and other properties of the dataset can impact on the overall performance of the framework.

A major issue with previous works was the use of n-gram model for feature extraction that have many deficiencies ,e.g, scaling, storage, no contextual information. One of the major issues in n-gram is the curse of dimensionality. The curse of dimensionality is that n-gram creates n dimensions for words. It means each word will be considered as a dimension in the model. As dimensions represent scaling, which means an extension in the feature set. In [17], it has been claimed that in many situations simple scaling will not help to make any significant progress. In existing research, no contextual information has been considered or extracted from the text. According to [23], “context-aware representations of words take into account both the ordering and identity of all the other words”. Many techniques have been proposed that uses contextual information in text classification and have outperformed n-gram model [17]. In the end, N-gram is limited to word level prediction and

because of n words it also gives importance to uncommon words, which are not needed necessarily. It is unable to extract sentence level features which can store more contextual information. Moreover, all the words have the same distance from each other in the n-gram model that ignores the importance of the specific words which can improve results in the classification. For example in blood request identification [20], some words were repeatedly used for blood request but due to the n-gram model, those words were assigned same probabilities like other common or uncommon words.

Chapter 3

Proposed Methodology

In this chapter, the scope and definition of the tweets are explained. After discussing the framework goal, data acquisition, data assessment, and data annotation are explained briefly. The next section is preprocessing in which all preprocessing steps are explained needed to clean data set for further processing. In the classification section, feature extraction models have been explained with their working structure. In the end, classifiers have been discussed used in this work.

3.1 Scope and Definition of Request Tweets

During critical situations and disasters millions of tweets may be generated using a hashtag. These tweets may or may not be relevant. If these tweets are relevant to the event, the question then becomes whether a tweet is a request or a normal statement such as not request. It is important to determine the criteria and definition of request, according to the Cambridge dictionary,"a request means asking for something or someone to do something in a polite or official way" [34]. However, in social media people do not follow the formal way of communication. According to [35], people make mistakes and use highly creative language while posting tweets. This shows that formal definition of request cannot be applied to tweets. As Twitter has limit of 280 characters, users do not care about grammar and correctness of the sentence[26]. People use predefined code and short words which makes tweets unstructured and

concise. Considering these factors, we have defined our generic scope for request identification, in which we not only consider formalities and politeness (,e.g., please, would you like to help, join us) but also include all those tweets that ask for help or service(s) or resource(s). These tweets contain tags used to mention other users, URLs which are used to donate, numbers used to text for donation. These request tweets can be interrogative or declarative. As tweets have been written informally, we will also consider tweets with short words ,e.g., hlp is used for help, plz/pls is used as please. There were also some confusing tweets which look like requests but they were suggestions and recommendations. As request tweets need an instant response, these tweets were excluded from request tweets so that the framework should identify actionable tweets.

3.2 Overview of Methodology

This framework aims to classify tweets into binary classes, such as, request or not request. The framework is built using different feature-based models and the combination of their features combined after extraction from the dataset. Features are extracted with their syntactic and semantic regularities. For contextual information and to avoid additional features, sentence level, and deep features extracted. As discussed earlier, tweets are posted using an informal language format. Considering misspelled and lack of grammar rules, it is very difficult to identify which tweets are requests and which are not. Sometimes two tweets having almost the same words with a little change in their order can be from different class. The model should be smart enough to classify these tweets in their respective class accurately. So the sole purpose of this work is the classification of tweets with better results as compared to previous work. The block diagram is given in figure 3.1.

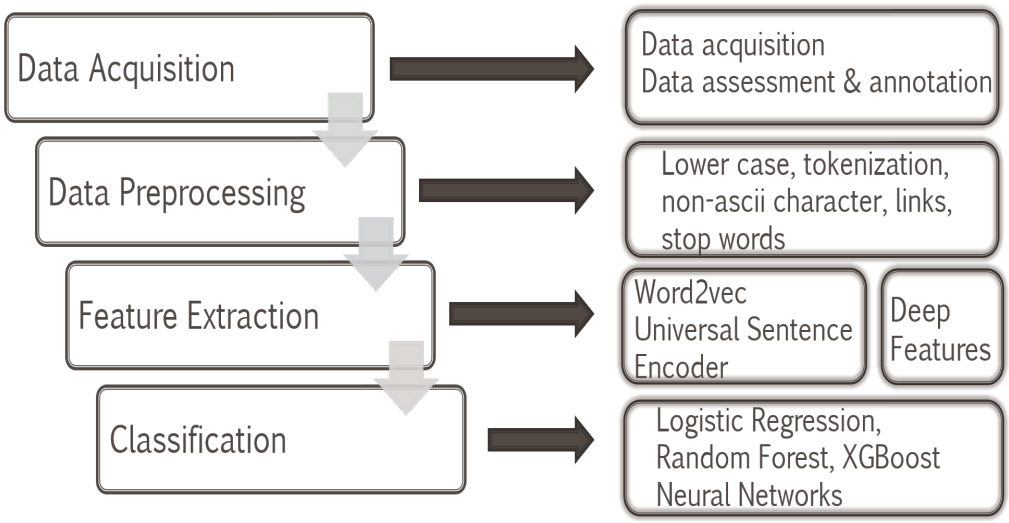

Figure 3.1: Block diagram of the proposed methodology

3.3 Data Acquisition

The dataset consists of three columns id, tweet, and label(tweet class). As our proposed model was purely based on text-based classification, it was useless to obtain additional information with the dataset. The dataset specification is given in Table 3.1.

Labels		
Request Tweets	Not Request Tweets	Total
528	2166	2744

Table 3.1: Dataset Specification before assessment

This dataset has two classes of tweets which are either requests or not request. The proposed system will learn and classify these tweets.

3.3.1 Data Assessment

All the tweets in the dataset were labeled as 'request' or 'not request'. While assessing tweets, it was observed that partial part of the dataset was incor-

rectly labeled. After a thorough assessment, around half of the tweets were identified which were tagged incorrectly. The details are given in Table 3.2.

Tweets Assessment		
Correctly Labelled	Incorrectly Labelled	Total
1349	1395	2744

Table 3.2: Dataset after assessment

The incorrectly labeled tweets were labeled again so that all tweets belong to the right class. The systematic way has been followed to label tweets.

3.3.2 Data Annotation

The annotation process can be performed automatically or manually. In [35], it has been claimed that automatic annotation is less precise as compared to manual annotation. As our work is about request classification, manual annotation is selected for the tweets. To carry on this process, many experts are requested and three of them showed interest. The tweets are given to each of them after collecting some necessary information. Details of the annotators are given in Table 3.3.

Gender	Age	Highest Degree	Employment Status
Male	30-39	PhD-Biomedical	40/week
Male	20-29	MS-Computer Science	1-39/week
Male	20-29	MS-Computer Science	40/week

Table 3.3: Expert Annotators Details

Total of 1395 tweets were annotated by each expert. All of them have a research background as well as industry experience. Annotation performed

on raw tweets without any preprocessing. An Excel sheet provided to all annotators which had all tweets with a dropout option of request and not request. The expert annotators read all these tweets one by one and selected one option from the drop-down menu. All the annotated files were stored and final results were obtained by implementing the voting technique [36]. After annotation which was done on half of the data, the dataset is combined for further processing.

3.4 Preprocessing

Preprocessing not only enhances the quality of the dataset but also reduces the inconsistency and noisiness of the data [37, 38]. It also reduces the size of data which optimize the resource usage and reduces the computation time. Following the preprocessing steps performed to clean the dataset.

3.4.1 Tokenization

Sentences have been broken down into words in this step. These words are tokens, used to create word embeddings. These tokens are used as input for further classification. This step is performed using python built-in library [39].

3.4.2 Removing non-ASCII characters

The non-ASCII characters are the symbols or signs which are not part of the formal English language. These characters are used in social media to make attractive and fancy posts but of no use in text processing and classification. So during preprocessing, all non-ASCII character are removed using regular expressions [1].

3.4.3 Removing Non-English Tweets

As this work is limited to globally known language English, all other tweets which are in other languages are removed ,e.g., "Como fue qe me enamoraste de esta forma???? Como lo isiste!!!!!Acaso me enamoraste x qe te gusto". All tweets which contain non-English words more than a specific limit or having full non-English tweets are eliminated from the dataset. Removing non-English tweets will not only enhances the performance of the framework but will also make it reliable.

3.4.4 Removing Stop Words

Words occurs frequently in the text are called stop words. These are just supporting words in sentences and do not have any semantic value. For example, prepositions, articles, conjunctions. According to [38, 40], these stop words do not provide any semantic and discriminative information which can help in text classification. A python library NLTK (Natural Language Toolkit) [48] contain a list of all stop words. This library is used in our framework to remove all stop words.

3.4.5 Removing tags and URLs

A lot of options are available in twitter while tweeting ,e.g., mentioning someone by using '@' symbol, adding URLs or website links, using some trending hashtags, etc. So to avoid all these useless junk text, Regular Expressions [41] are used to remove it in Python. In Figure 3.2, summary of Regular Expression Patterns is given.

3.4.6 Removing Duplicate Tweets

When tweets were extracting from Twitter using different hashtags, some users used more than one hashtags in their tweets. In result, a lot of duplicate tweets were extracted. These duplicate tweets can create bias in text classification. To avoid this, it was necessary to remove duplicate tweets. And after some

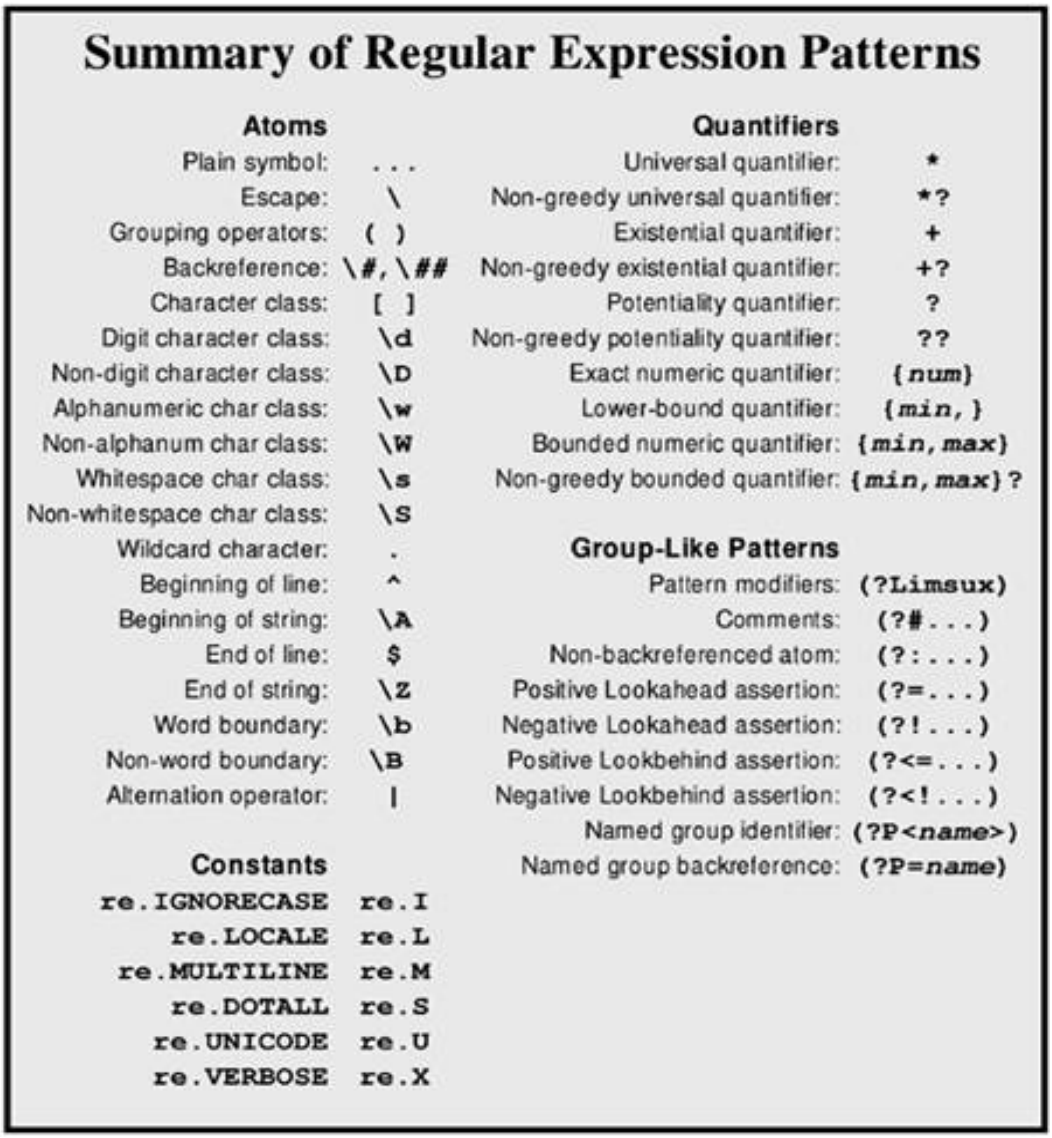

Figure 3.2: Summary of Regular Expression Patterns [1]

preprocessing tweets, it was seen that some tweets became duplicate. All duplicate tweets are removed using set property in Python and unique tweets are stored [42].

3.5 Classification

Classification is a problem in which an algorithm tries to predict a class for targeted data. Classification is based on two parts. First, extract features from the dataset. Secondly, that feature set is used as input to the classifiers

with a test data which is classified in targeted classes.

3.5.1 Feature Extraction

Dataset, in this context 'tweet', is written in a vague manner and contains a lot of redundant, extra and useless data. Features are meaningful, concise, well-arranged attributes that contain a maximum representation of the data. A well-defined feature set can impact the results significantly. In this work, we have selected two feature extraction models which are the state of the art nowadays in text classification. One model extracts word-level semantic and syntactic features while the other model extracts sentence-level contextual features. Both models will be explained briefly in upcoming sections.

3.5.1.1 Word2vec Model

The word2vec[22] is a feature extraction model which creates text features considering semantic and syntactic regularities at the word level. This model is divided into two main steps. In the first steps, a dictionary is created from the given dataset and stored as a matrix for further use. The matrix provides distributed representation in a vector space of different words to help learning algorithms or classifiers. These word representations are added using neural networks and are very surprising as they explicitly consider and encode semantic, syntactic and linguistic regularities. The words with the same context are grouped together to make the better prediction. The input layer takes a corpus of words and the output layer produced a set of vectors which contains numerical values. The similarities it detects between words are based on complex mathematical calculations. The graphical representation of the distributed words based on numerical values is given in figure 3.3. In the next step, these stored values are assigned to the dataset which will replace all words with their numerical values obtained from the dictionary created by the word2vec model. These new numerical values are called word embeddings and used as classifierâĂŹs input for classification.

Figure 3.3: Distributed representation of similar words

3.5.1.2 Universal Sentence Encoder

The Universal Sentence Encoder[23] is a feature extraction model that creates text features considering semantic and syntactic regularities at the sentence level. It provides two variants for embedding vectors at the sentence level. One is based on transformer architecture proposed in [43] while the other is based on Deep Averaging network proposed in [44]. But only one was avail-

able at the time of experiments. So deep averaging network model is used. In Deep Averaging Network, words from the sentence are averaged together with their bigrams and then sentence level embeddings are produced by neural networks. These calculated sentence level embeddings are converted to a fixed length matrix which contains all contextual information and relationship among the words. It first converted the sentence into lower case and then converted the whole sentence into a token. Then this token is taken as input and 512 dimension matrix is produced having all contextual value of the words that were in the input. This output is also called sentence embedding vector. Unlike Transformer Architecture that requires high resources of computations, Deep Averaging Network uses resources in an efficient way with a slight decrease in accuracy.

3.5.1.3 Hybrid Features

Deep features are extracted with the combination of word-level embeddings and sentence level embeddings. First, a dictionary was created from the given dataset and stored its matrix using word2vec model. This dictionary was created by the input of the dataset. It stored distributed representation of the similar words used in the dataset. It created a contextual relation between these words and stored their numerical values in matrices. Then, assigned these stored values to the dataset which replaced all words with their numerical values obtained from the dictionary created by the word2vec model. These numerical values which are in matrices are stored as feature set for creating deep features. In the next step, words from the sentence are averaged together with their bigrams and then sentence level embeddings are produced by neural networks. These calculated sentence level embeddings are converted to a fixed length matrix which contains all contextual information and relationship among the words. It first converted the sentence into lower case and then converted the whole sentence into a token. This token was taken as input and 512 dimension matrix was produced having all contextual value among words that were in the input. This output is also called sentence embedding vector. The matrix is obtained by multi-task learning where a single model is further

become the input of multiple calculation models. After successful creation of sentence-level embedding, both word level feature set from word2vec and sentence level feature from sentence encoder are concatenated. This concatenation then produced deep feature as input for classifiers to do classification.

3.5.2 Classifiers

A list of classifiers is available and all of them have pros and cons. Some classifiers perform better while others do not, it depends on the situation. Some classifiers are linear models while others are ensemble models. Linear models are a single algorithm that takes input and returns classified output. While ensemble models are a combination of different linear algorithm which means they can produce optimized results as compare to linear algorithms. This work has implemented both types of models to compare and find better results.

These classification models are optimized using parametric tuning. The objective was to find optimized parameters which can give maximum performance. There were two ways to do parametric tuning. Manually parametric tuning or by search algorithm which automatically tries all possible parameters with all possible combinations to find the best parameters in the specific problem. While in manual parametric tuning, different available options are tried one by one to find the most suitable parameters that were producing the best results. Manual parametric tuning was very hectic as changes are made again and again while on the other hand search algorithm is also applied, it took some extra time but it also gave best parameters of the classifier for the current problem. However, final best results were achieved by manual parametric tuning. Following four classifiers were used to classify request tweets in our framework on all three feature sets in Table 3.4.

Classifiers	Feature set
Logistic Regression	Word2vec
Random Forest	Universal Sentence Encoder
XgBoost	Hybrid Features
Neural Networks	

Table 3.4: Classifiers and feature set

3.5.2.1 Logistic Regression

Logistic Regression is one of the linear classifiers that is used to set a baseline and gave an idea about the dataset [45]. The reason for using Logistic Regression is, it understands the relationship between feature set and the categories. Basically logistic regression created an understanding between dependent variables and independent variables using a complex mathematical function. The function used to understand the relation between independent and dependent is called the SoftMax function. The output of the SoftMax function always remains between 0 and 1. The results of this classifiers will be explained in the next chapter of results and evaluation.

3.5.2.2 Random Forest

Random forest has many calculation trees [46]. Each tree calculated and predicted the target class. The trees basically identified some features from the dataset and tried to predict a class based on these features. Due to flexibility in feature selection, overfitting problem did not occur. Another benefit of the Random Forest was, it helped in feature engineering and gave important features that enhanced results and performance.

3.5.2.3 XgBoost

XgBoost is the extension of a gradient boosting tree algorithm [17]. The reason to use this classifier was, it has very good execution speed and improved the model performance. In addition, its resource management was very sys-

tematic and it has a wide variety of parameters for tuning and cross-validation. It started with a weak prediction and then by selecting better features and parameters, its went exponentially towards best results. XgBoost is based on tree ensemble which is a set of classification and regression trees.

3.5.2.4 Neural Networks

Neural Network is a state of the art classifier nowadays. The purpose of using Neural Networks was, it created a better understanding of deep features in which world level and sentence level features were used [47]. As it resembles human brains and has many neuron layers with weighted connections, it learned again and again the context of the tweets and gave best classification results as compared to other three classifiers.

Chapter 4

Results and Evaluation

First, evaluation measures are discussed in this chapter that will be used through whole chapter for results and experiments evaluation. These evaluation measures are Accuracy, F1 score, Precision, Recall, and two different cross-validation techniques. After discussing evaluation measures, different feature extraction models are experimented. It started with Word2vec model then Universal Sentence Encoder and at the end, combination of both models were used. Each model is tested initially with four classifiers which are Logistic Regression, Random Forest, XgBoost, and Neural Networks. Different evaluation metrics are calculated with best classifier and selected as best classifier. After all evaluation, cross-validation is tested and final results are discussed.

4.1 Evaluation Metrics

This work is based on machine learning and data science. Some evaluation metrics are used as standard for this type of work [48]. These metrics are Precision, Recall, Accuracy, and F1 score.

4.1.1 Precision

Precision is also a performance measure which checks from all classes how much a classifier predicted correctly. The highest precision means classifier

has performed well and gave better results [49].

4.1.2 Recall

Recall is a performance measure which checks from all positive how much correctly predicted by the classifier. The higher recall show better performance of the classifier [49].

4.1.3 Accuracy

Accuracy measure calculated a value which shows the closeness to the actual value. The higher accuracy means the calculated value is closer to the actual value and results are better [48].

4.1.4 F1 Score

Sometimes precision and recall have very distant values from each other and it seems difficult to evaluate and comparison. To make a proper comparison in which both precision and recall are used, F1 score is used which is the harmonic mean of the precision and recall. The higher F1 score mean better overall performance [48].

4.1.5 Hold Out Validation

This is a simple validation technique in which data is split in two sets, one is called training the other is called testing set [50]. These two sets can be split into any ratio depending on the nature of the work e.g. 70:30, 60:40, or it could be 80:20. Our work has been produced on 70:30 ratio. A function simply splits data randomly into two sets from any point, one set is selected as training set which is generally larger than the other set called test set .

4.1.6 Cross Validation

Cross validation is also called K-fold cross validation. In this validation technique, k fold is defined and data divided into those k groups [50]. Then one group is selected as a test set while rest are combined and selected as a training set. This process continues until all groups are used as test data. At the end, average is taken of the all results.

4.1.7 Data Specification

This work has been carried out on the dataset of tweets posted during Hurricane disaster. The specification of the dataset are given in Table 3.1.

4.2 Evaluation with Word2vec features

This model has been divided into two main steps. In the first steps, it creates a dictionary from the given dataset and stores its matrix for further use. This dictionary is not based on general English words but created by the input of the dataset. It stores distributed representation of the similar words used in the dataset [22]. It creates a contextual relation between these words and stores their numerical values in matrices. The next step is to assign these stored values to the dataset which will replace all words with their numerical values obtained from the dictionary created by the word2vec model. These new numerical values are called word embeddings and used by classifiers for classification. After extracting features, initially four classifiers have been tested on these features. The results are given in Table 4.1.

This experiment has been carried out with Hold-out validation. From the results as we can see Logistic Regression and Neural Networks has given same results which are low as compared to other two classifiers. The reason of the low results could be that these two models are linear model and it may be unable to create understanding with the features of the data. Another reason

Classifier	Accuracy
Logistic Regression	79.24
Random Forest	85.19
XgBoost	84.22
Neural Networks	79.24

Table 4.1: Classifiers with Accuracy Results

behind low performance of these to classifiers could be the matrix‘s dimensions. In word2vec model, 100 dimensional feature matrix has been created against each word which may provide less information as these two classifiers are not good with less data. On the other hand, Random Forest and XgBoost gave better results with such small embeddings. These two classifiers are very good in feature understanding and identified best features using tree structure [64]. Moreover, these features contains syntactic and semantic information which has also provided better understanding with tweets and the results are much better than the proposed baseline model. We will see more evaluation matric with Random Forest and XgBoost. In figure 4.1, a chart is showing the performance of the Random Forest and XgBoost with classified tweets.

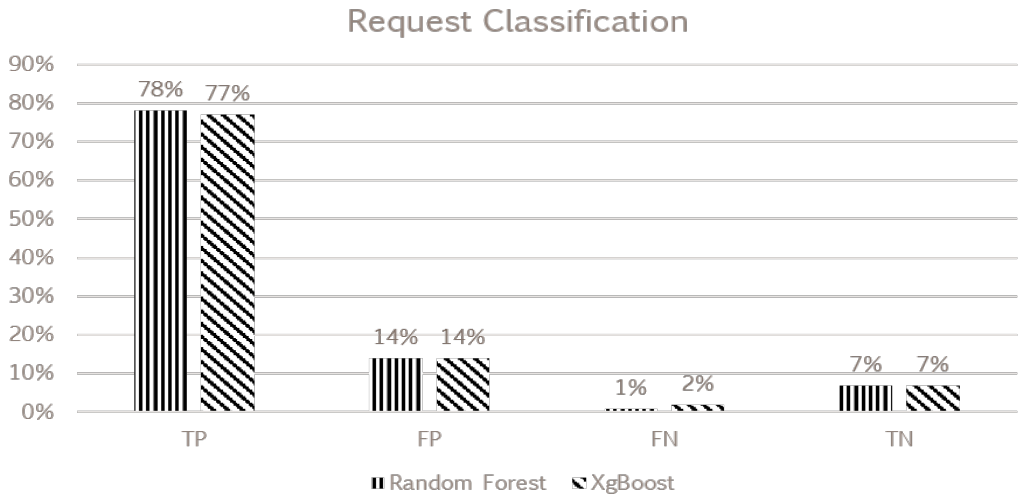

Figure 4.1: Chart of Classified Tweets

As it can be seen in figure 4.1, that true positive and true negative are correct

classification while false positive and false negative are the errors or wrongly classified tweets by the classifiers. True positive and false negative belongs to the same class 'not request'. The remaining two TN and FP belongs to the same class 'request' Now for the evaluation and comparison, other evaluation measures are calculated in Table 4.2 to discuss detailed performance of the classifiers.

Measures	Random Forest(%)	XgBoost(%)
Accuracy	85.10	84.20
F1	91.30	90.70
Precision	98.50	97.50
Recall	85.10	84.80

Table 4.2: Random Forest and XgBoost results with Hold-out

In Table 4.2, Hold-out validation has been used. The Random Forest has given best results as compared to XgBoost and other two classifiers using word2vecc model. Accuracy, F1 score, Precision, and Recall have been calculated for Random Forest and XgBoost classifiers. Accuracy shows overall performance which is 85% and 84% respectively. Precision is 98% and 97% which is highest figure in this table shows that how much tweets correctly classified from all classified tweets. In Recall which is 85% and 84% shows how much tweets are correctly classified from all correct tweets. F1 score which shows the overall performance for Precision and Recall is 91% and 90%, that means overall classifiers has given better results.

Cross validation with 10-fold has also been experimented with Random Forest and XgBoost as these classifiers have produced better results. In this validation method, data is divided into 10 folds. Nine folds are combined and used for training while other one is tested with classifier. This method continues

Measures	Random Forest(%)	XgBoost(%)
Accuracy	84.84	84.10
F1	50.38	49.65
Precision	79.43	76.61
Recall	37.25	37.25

Table 4.3: Random Forest and XgBoost results with 10-fold

until all folds are tested. In Table 4.3, accuracy is 84% for both classifiers. But as dataset was imbalanced, Recall did not produced better results. Due to small dataset, precision is also low. F1- score is also not satisfactory. This shows, due to data limitation cross validation did not produce expected results as compared to Hold-out validation.

4.3 Evaluation with Universal Sentence Encoder

In this model, sentence level embeddings have been produced by neural networks. These calculated sentence level embeddings have been converted to a fixed length matrix which contains all contextual information and relationship among the words. It first converted the sentence into lower case and then converted the whole sentence into a token. This token has been used as input and 512 dimension matrix has been produced having all contextual value among words that were in the input. This output is also called sentence embedding vector [23]. The matrix has been obtained by multi-task learning where a single model is further become the input of multiple calculation models. After extracting features, initially four classifiers have been tested on these features. The results are given in Table 4.4.

This experiment has been carried out with Hold-out validation. From the results, we can see Random Forest has given results which are lowest as compared to other three classifiers. The reason of the low results could be that features are increased as compared to the model used earlier. Another rea-

Classifier	Accuracy
Logistic Regression	87.01
Random Forest	84.46
XgBoost	86.89
Neural Networks	86.16

Table 4.4: Classifiers with Accuracy Results

son behind low performance of these to classifiers could be the matrixâĂŹs dimensions. In this model, 512 dimensional feature matrix has been created against each word which may have provided some confusing information as this classifier is highly dependent to the amount of features [46]. On the other hand, Logistic Regression, Neural Networks, and XgBoost has given better results with such wide embeddings. Because these three classifiers are very good in feature understanding and can identify best features. Moreover, these features contains syntactic and semantic information and word‘s relationship at sentence level which has also provided better understanding with tweets and the results are much better than the proposed baseline. As Logistic Regression and Neural Networks have given best results compared to other two classifiers. We will see more evaluation metrics with Logistic Regression. In figure 4.2, a chart is showing the performance of the Logistic Regression and Neural Networks with classified tweets.

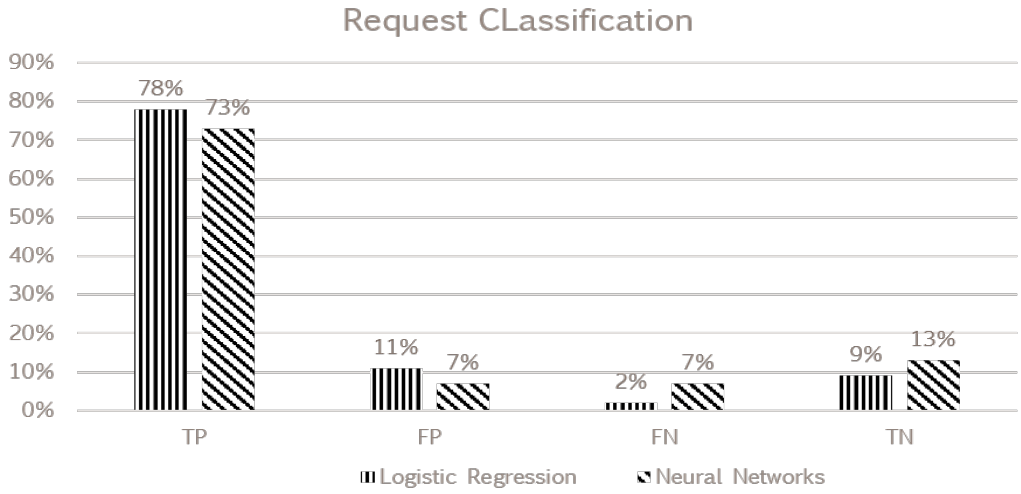

Figure 4.2: Chart of Classified Tweets

As it can be seen in figure 4.2, that true positive and true negative are correct classification while false positive and false negative are the errors or wrongly classified tweets by the classifiers. True positive and false negative belongs to the same class 'not request'. The remaining two, TN and FP belongs to the same class 'request'. Now for the evaluation and comparison, other evaluation measures have been calculated in Table 4.5 to discuss detailed performance of the classifiers.

Measures	Logistic Regression(%)	Neural Networks(%)
Accuracy	87.01	86.20
F1	92.30	91.30
Precision	98.00	91.90
Recall	87.20	90.80

Table 4.5: Logistic Regression and Neural Networks results with Hold-out

In Table 4.5, Hold-out validation has been used. The Logistic Regression has given best results as compared to Neural Networks and other two classifiers using universal sentence encoder. Accuracy, F1 score, Precision, and Recall have been calculated for Logistic Regression and Neural Networks classifiers. Accuracy shows overall performance which is 87% and 86% respectively. Precision of Logistic Regression is 98% which is much better than Neural Networks which is 91%. Precision shows that how much tweets correctly classified from all classified tweets. However in Recall, Neural Networks has achieved 90% as compared to Logistic Regression which is 87%. Recall shows how much tweets are correctly classified from all correct tweets. F1 score which shows the overall performance for Precision and Recall is 92% and 91% for Logistic Regression and Neural Networks respectively, that means overall classifiers has given better results.

Cross validation with 10-fold has also been experimented with Random Forest

Measures	Logistic Regression(%)	Neural Networks(%)
Accuracy	86.92	86.66
F1	62.08	63.60
Precision	78.56	74.31
Recall	51.50	56.00

Table 4.6: Logistic Regression and Neural Networks with 10-fold

and XgBoost as these classifiers have produced better results. In this validation method, data is divided into 10 folds. Nine folds are combined and used for training while other one is tested with classifier. This method continues until all folds are tested. In Table 4.6, accuracy is 84% for both classifiers. But as dataset was imbalanced, Recall did not produced better results. Due to small dataset, precision is also low. F1- score is also not satisfactory. This shows, due to data limitation cross validation did not produce expected results as compared to Hold-out validation.

4.4 Evaluation with Hybrid Features

This is the last and final model of features in this work. In hybrid features, word-level and sentence-level features are combined. These features are the best features as they are extracted from top feature models. One model word2vec contains syntactic and semantic regularities while other contains sentence level features that represent contextual information [22, 23]. By combining these models, the performance of the classifiers will be better especially the one which performs better with more features. This model contains 612 dimensional embedding matrix for each word and initially tested on four classifier. The one with best result has been chosen as best classifier and other evaluation metrics have been calculated for it. The results of four classifiers are given in Table 4.7 tested on hybrid features.

The classifiers are tested using hold-out validation and results are obtained.

Classifier	Accuracy
Logistic Regression	86.89
Random Forest	84.58
XgBoost	86.77
Neural Networks	89.07

Table 4.7: Classifiers with Accuracy Results

It can be clearly seen that Neural Networks has outperformed other three and gave the best results so for. The reason behind other three classifiers, as they did not produced better results could be the complexity of the features. Logistic Regression results are 2nd highest in this last model because the SoftMax function worked as expected but was unable to understand the depth of the features [45]. Random Forest results were the lowest, the reason could be misunderstanding of features as features were extracted from different models. The XgBoost was at 3rd number in performance due to its complex tree structure in which multiple trees extract different features and based on these features, they tried to classify tweets into their respective class [17]. As neural networks is more intelligent classifier than the other three and able to understand both types of features which were extracted at word level and at sentence level. It produced the best results and improved the results significantly as compared to baseline model. Another reason behind Neural Network‘s performance is, it had two hidden layer which makes it an deep learning algorithm. The deep learning algorithm usually performs well when the features are more complex. In figure 4.3, a chart is showing the performance of the Neural Networks with classified tweets.

As it can be seen in figure 4.3, that true positive and true negative are correct classification while false positive and false negative are the errors or wrongly classified tweets by the classifier Neural Networks. True positive and false negative belongs to the same class ‘not request’. The remaining two TN and FP belongs to the same class ‘request’. Now for the record and comparison, other evaluation measures are calculated in Table 4.8 which shows overall performance of the classifier.

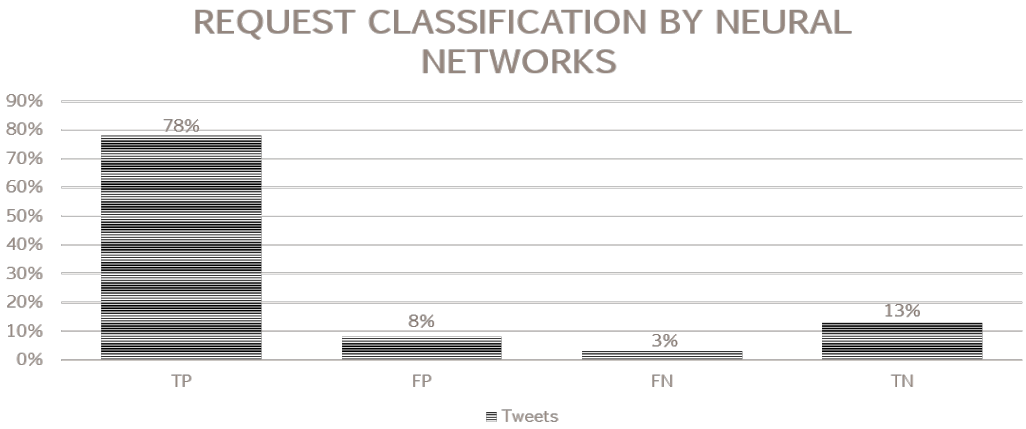

Figure 4.3: Chart of Classified Tweets

Measures	Neural Networks(%)
Accuracy	89.07
F1	93.30
Precision	96.00
Recall	90.70

Table 4.8: Neural Networks results with Hold-out

In Table 4.8, Hold-out validation has been used. The Neural Networks has given best results as compared to other three classifiers using hybrid features. Accuracy, F1 score, Precision, and Recall have been calculated for Neural Networks. Accuracy shows overall performance which is 89%. Precision is 96%. Precision shows that how much tweets correctly classified from all classified tweets. In Recall, Neural Networks has achieved 90% which is best so far in this work. Recall shows how much tweets are correctly classified from all correct tweets. F1 score which shows the overall performance for Precision and Recall is 93%, that means overall classifiers has given better results.

Cross validation with 10-fold has also been experimented with Neural Networks as this classifiers has produced better results. In this validation method,

Measures	Neural Networks(%)
Accuracy	86.40
F1	67.57
Precision	71.48
Recall	64.72

Table 4.9: Neural Networks results with 10-fold

data is divided into 10 folds. Nine folds are combined and used for training while other one is tested with classifier. This method continues until all folds are tested. In Table 4.9, accuracy is 86%. But as dataset was imbalanced, Recall did not produced better results. Due to small dataset, precision is also low. F1- score is also not satisfactory. This shows, due to data limitation cross validation did not produce expected results as compared to Hold-out validation. However, hybrid features have given best results as compared to word2vec and universal sentence encoder model.

4.5 Results and Discussion

This research work has given satisfactory and better results than the baseline models. In Table 4.10 results are compared with old dataset. Due to incorrect labelling our framework was not able to create contextual understanding, however it gave improved results in accuracy recall and F1 than the baseline model and [21]. In Table 4.11, a result comparison is made with previous work on new dataset. Our model has given best results in both Hold-out validation and cross validation. From Hold-out validation, the highest evaluation measure is Precision with 96% results. Recall is 90%. While two standard evaluation measures, Accuracy and F1 score used for comparison have results of 89% and 93% respectively. Overall our system has given best results with new state of the art feature extraction models and machine and deep learning classifiers. Three different feature extraction models are used with four same classifiers for each feature model. However, in cross validation results were not as good as compared to Hold-out validation. The reason behind it is im-

balanced and small dataset. The last model hybrid features in Table 4.8 with the classifier Neural Networks has produced best results for both Hold-out and cross validation in this work.

Method	Validation	Accuracy(%)	Precision(%)	Recall(%)	F1(%)
This Framework	10-fold	83.33	78.9	85.2	82.00
Irfanullah 2018	10-fold	82.37	83.37	82.38	82.35
Nazer 2016 [21]	10-fold	80.28	80.37	80.28	80.28

Table 4.10: Comparison of results with actual data

Method	Validation	Accuracy(%)	Precision(%)	Recall(%)	F1(%)
This Framework	Hold-out	89.10	96.00	90.70	93.30
	10-fold	86.40	67.57	71.48	64.72
Irfanullah 2018	10-fold	54.92	54.70	54.92	53.51

Table 4.11: Comparison of results with assessed data

These results are not just based on feature extraction models and classifiers. But data assessment and data annotation has been followed systematically. Preprocessing has also played vital role in the performance of the system which includes, tokenization, cleaning tweets, removing duplicates, and lower case conversion of text.

Chapter 5

Conclusion and Future Work

This chapter contains brief explanation about the work, that has been accomplished and how the research objectives were achieved. Then a brief discussion of the significance of this research has been explained. In the end, limitations and future work has been indicated further improvements.

5.1 Conclusion

Tweets classification has become popular over time. People use Twitter as more than just a social application. Researchers have identified many research problems that can support in human lifestyle. Question-answers, opinion mining, surveys, and request identification are popular research areas on Twitter. Specially request tweets classification is an important research problem, where tweets are posted in critical situation or in the time of need. A lot of research efforts have been made for request classification. A major drawback in these research has been the use of n-gram model for features extraction. This approach does not represent contextual information of the text. As size of data increases, these techniques suffer from curse of dimensionality. In our work, use of more reliable and recently introduced techniques has been proposed. These techniques not only use contextual information of the text but also avoid curse of dimensionality. Moreover, sentence level features have been used for request classification. By merging features from two different models, hybrid features have been extracted that produced better results as

compared to baseline models. For evaluation and verification, hold-out and cross-validation approaches have been used. Evaluation metrics, precision, recall, accuracy, and F1 measures have been used for comparison and evaluation. In this work, 89% accuracy has been achieved using hybrid features and neural network classifier for request classification. Other measures like, precision, recall, and F1 score were 90%, 96%, 93% respectively. These results were obtained using hold-out validation.

5.2 Contribution of Research

This research work is an important step forward in the journey of research related to request identification. New proposed techniques were used to address shortcoming of previous work. The major contribution was the use of contextual information which helped in getting better performance. Sentence level feature extraction was another achievement in this research. Extra features, like metadata, were avoided which creates overhead. The curse of dimensionality during feature selection has also been avoided. Two different models for feature extraction were merged together which have produced better results. Experiments have been performed using advanced classifiers. Thorough hyperparametric tuning was applied on classifiers to get maximum output. In conclusion, a better system has been developed that can classify request tweets successfully with improved performance as compared to previous frameworks.

5.3 Limitations and Future work

The system has produced better results, however the dataset was relatively small. The system performance also depends on the ratio of classes in a dataset. In this work, tweets were not equal for both classes and dataset was imbalanced. Two different variants were proposed for sentence level feature extraction. One was based on deep averaging networks and implemented in this work , this variant uses efficient resources with slightly low results. Other is based on transformer architecture and requires high resources. It was

unavailable at the time of experiments. So this variant is not implemented in this work. In future, dataset can be improved, request tweets may be added to make system more efficient. Request classification is a broad problem and different types of requests are made during disasters and in critical situations. It is possible to combine request classification data from different aspects or from situations. This may not only give better results but also help to create a more reliable system. Another addition to this work could be balancing the dataset. It might give better results with a balanced dataset. Another task that can be accomplished in future is implementing the other variants of the model used for sentence level feature extraction. This model was not implemented due to unavailability and high resource requirement.

References

[1] David Mertz. *Text Processing with Python*. Addison-Wesley Longman Publishing Co., Inc., 2003.

[2] Spence P.R. Van Der Heide Westerman, D. A social network as information: The effect of system generated reports of connectedness on credibility on twitter. *Computers in Human Behavior*, 28:199–206, 2012.

[3] https://en.wikipedia.org/wiki/Twitter. Twitter. 2006.

[4] https://www.alexa.com/topsites. Top sites. 2018.

[5] J. Boase M. Naaman and C. H. Lai. Is it really about me?: message content in social awareness streams. *Computer supported cooperative work*, 28:189–192, 2010.

[6] C. Honey and S. C. Herring. Beyond microblogging: Conversation and collaboration via twitter. *System Sciences*, 42:1–10, 2009.

[7] D. Zhao and M. B. Rosson. How and why people twitter: The role that micro-blogging plays in informal communication at work. *International Conference on Supporting Group Work*, pages 243–252, 2009.

[8] J. Teevan M. R. Morris and K. Panovich. What do people ask their social networks, and why?: a survey study of status message qa behavior. *Human factors in computing systems*, pages 1739–1748, 2010.

[9] R. Gazan. Social qa. *Journal of the Association for Information Science Technology*, 62:2301–2312, 2011.

[10] S. Joshi. Twitter, facebook and google activate features to help people affected by chennai floods. 2017.

[11] V. Goel S. Ember. As paris terror attacks unfolded, social media tools offered help in crisis. 2015.

[12] S. E. Vieweg. Situational awareness in mass emergency: A behavioral and linguistic analysis of microblogged communications,. *Doctoral dissertation*, 2012.

[13] Paul M. J. Palmer M. Palen L. Anderson K Stowe, K. Identifying and categorizing disaster-related tweets. *Natural Language Processing for Social Media*, pages 1–6, 2016.

[14] A. Acar and Y. Muraki. Twitter for crisis communication: lessons learned from japan's tsunami disaster. *Web Based Communities*, 7:392–402, 2011.

[15] D. Murthy and S. A. Longwell. Twitter and disasters: The uses of twitter during the 2010 pakistan floods. *Information, Communication Society*, 16:837–855, 2013.

[16] Daniel Jurafsky James H. Martin. Speech and language processing. 2014.

[17] Chen K. Corrado G. Dean J Mikolov, T. Efficient estimation of word representations in vector space.

[18] Manaf N. A. A. Iannone L. Stevens Mikroyannidi, E. Analysing syntactic regularities in ontologies. 849, 2012.

[19] Richmond H Thomason. What is semantics. *Education*, 2012.

[20] Ayyar M. Chopra S. Shahid S. Mehnaz L. Shah R. Mathur, P. Identification of emergency blood donation request on twitter. *Social Media Mining for Health Applications Workshop Shared Task*, pages 27–31, 2018.

[21] Harsh Dani Tempe Tahora H. Nazer, Fred Morstatter. Finding requests in social media for disaster relief. *ASONAM)*, 2016.

[22] Sutskever I. Chen K. Corrado G. S. Dean J Mikolov, T. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, pages 3111–3119, 2013.

[23] Yang Y. Kong S. Y. Hua N. Limtiaco N. John R. S. Sung Y. H Cer, D. Universal sentence encoder. 2018.

[24] B. Diri Z. B. Ozger and C. Girgin. Question identification on turkish tweets," in innovations in intelligent systems and applications (inista) proceedings. *International Symposium*, pages 126–130, 2014.

[25] M. R. Lyu I. King Li, X. Si and E. Y. Chang. Question identification on twitter. *International Conference on Information and Knowledge Management*, 20:2477–2480, 2011.

[26] K. Dent and S. Paul. Through the twitter glass: detecting questions in micro-text. *Conference on Analyzing Microtext*, pages 8–13, 2011.

[27] T. Elsayed M. Hasanain and W. Magdy. Identification of answer-seeking questions in arabic microblogs. *Information and Knowledge Management*, 23:1839–1842, 2014.

[28] M. Efron and M. Winget. Questions are content: a taxonomy of questions in a microblogging environment. *Information Science and Technology*, 47: 1–10, 2010.

[29] J. Logie F. M. Harper, J. Weinberg and J. A. Konstan. Question types in social qa sites. 15, 2010.

[30] T. Ragavan N. Prasath AB. Gokulakrishnan, P. Priyanthan and A. Perera. Opinion mining and sentiment analysis on a twitter data stream. *Advances in ICT for emerging regions (ICTer)*, pages 182–188, 2012.

[31] Roy D. Vosoughi, S. Tweet acts: A speech act classifier for twitter. *Conference on Web and Social Media.*

[32] J. Lucas P. Meier M. Imran, C. Castillo and S. Vieweg. Aidr: Artificial intelligence for disaster response. *World Wide Web*, 23:159–162, 2014.

[33] Li Y. Reiss F. R Chiticariu, L. Rule-based information extraction is dead! long live rule-based information extraction systems. *empirical methods in natural language processing*, pages 827–832, 2013.

[34] Cambridge Dictionary. Cambridge advanced learnerâĂŹs dictionary. *Recuperado de: https://dictionary. cambridge. org/es/diccionario/ingles/blended-learning*, 2008.

[35] Baycroft Petrillo, M. Introduction to manual annotation. *Fairview Research*, 2010.

[36] Stumpf A. Gancarski P. Lampert, T. A. An empirical study into annotator agreement, ground truth estimation, and algorithm evaluation. *Transactions on Image Processing*, 25(6):2557–2572, 2016.

[37] G. P. Varma I. Hemalatha and A. Govardhan. Preprocessing the informal text for efficient sentiment analysis. *Emerging Trends Technology in Computer Scienceb*, 1:58–61, 2012.

[38] J. Lucas P. Meier M. Imran, C. Castillo and S. Vieweg. The impact of preprocessing on text classification," information processing management. *Information Processing Management*, 50:104–112, 2014.

[39] https://docs.python.org/3/library/tokenize.html. Tokenizer for python source. 2018.

[40] J. PomikÃąlek and R. Rehurek. The influence of preprocessing parameters on text categorization. *International Journal of Applied Science*, 1: 430–434, 2007.

[41] https://docs.python.org/3/library/re.html. Regular expression. *Library*, 2018.

[42] https://docs.python.org/2/library/string.html. Common string operations. 2018.

[43] Niki Parmar Jakob Uszkoreit Llion Jones Aidan N Gomez Kaiser Ashish Vaswani, Noam Shazeer and Illia Polosukhin. Attention is all you need. *NIPS*, 2017.

[44] Jordan Boyd-Graber Mohit Iyyer, Varun Manjunatha and Hal Daum III. Deep unordered composition rivals syntactic methods for text classification. *IJCNLP*, 2015.

[45] Dietz K. Gail-M. Klein M. Klein M. Kleinbaum, D. G. Logistic regression. *New York: Springer-Verlag*, 2002.

[46] L Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[47] S. S. Haykin. Neural networks and learning machines/simon haykin. *Prentice Hall*, 2009.

[48] http://cs229.stanford.edu/section/evaluation$_m$*etrics.pdf*.*Evaluationmetrics*.2018.

[49] Goadrich M Davis, J. The relationship between precision-recall and roc curves. *In Proceedings of the 23rd international conference on Machine learning*, pages 233–240, 2006.

[50] Shukla S Yadav, S. Analysis of k-fold cross-validation over hold-out validation on colossal datasets for quality classification. *IACC*, pages 78–83, 2016.