Usable Amharic Text Corpus for Natural Language Processing Applications

Michael Melese Woldeyohannis, Million Meshesha
Addis Ababa University, Addis Ababa, Ethiopia

Abstract

In this paper, we describe the preparation of a usable Amharic text corpus for different Natural Language Processing (NLP) applications. Natural language applications, such as document classification, topic modeling, machine translation, speech recognition, and others, suffer greatly from a lack of digital resources. This is especially true for Amharic, a resource-constrained, morphologically rich, and complex language. In response to this, a total of 67,739 Amharic news documents consisting of 8 different categories from online sources are collected. The collected corpus passes through a number of pre-processing steps including; data cleaning, text normalization and punctuation correction. To validate the usability of the collected corpora from different domains, a baseline document classification experiment was conducted. Experimental results show that, 84.53% accuracy is registered using deep learning in the absence of linguistic information. Finding indicated that it is possible to use the prepared corpora for different natural language applications in the absence of linguistic resources such as stemmer and dictionary despite the complexity of Amharic language. We are further working towards Amharic news document classification by incorporating a linguistic independent stop-word detection, stemming and unsupervised morphological segmentation of Amharic documents.

Keywords: Amharic, Semitic, Under-resourced, Text Corpus, Document Classification

1. Introduction

Information is exchanged between different people in the form of verbal, nonverbal, written and visual communication. Text (written) based information and knowledge sharing is the most widely used means of communication compared to other modes of communication in the web and workplace (Park, Chung, and Lee, 2012). Information is exchanged through web, news, books, pamphlets, blogs, letters, Emails and memos.

The consumption of news in particular has changed drastically in the period from the pre-internet to the internet when people tune into events happening around the world through different channels (Leiner et al., 2009). Among these channels, news, social media and blogs are the major means for getting information from the internet in various forms (Li, 2013). The advancement of the internet as a means of communication has led to an ever-increasing demand for Natural Language Processing (NLP) which contributes to the provision of information through a number of applications. These applications include machine translation, speech recognition, text summarization, information retrieval, information extraction, topic modeling, sentiment analysis and text classification among others (Jurafsky and Martin, 2008). The last-mentioned, text classification is method of classifying text involves putting it into well-organized groupings. Text classifiers can manually or automatically assess text using NLP and categorize it according to its content (Sammut and Webb, 2011). The classification might be at the level of document, paragraph, sentence or sub-sentence level depending on the need.

According to Eberhard and Fennig (2020), there are around 7,151 living languages in the world, of which the majority have limited lack of availability of digital resources such as text corpora, tools and experts for different NLP tasks and this is especially true for Ethiopian languages. Ethiopia is a multilingual and multi-ethnic country ranked 48th in the world for language diversity contributing around 90 languages to the world. In addition to the diversity of the languages, different NLP applications require enormous amount of data and they are resource-intensive.

The challenges involved in developing NLP applications in Ethiopia are subject to technological and linguistic factors. The technological factor that affects natural language application depends on the resource requirement of classical and ensemble machine learning, and on deep learning techniques. Classical machine learning is used to predict new observations, or determine the output of new input, and ensemble machine learning methods combine the prediction from two or more models and suggest the best solution. These machine learning algorithms are not preferred for solving complex tasks that require huge amounts of data. On the other hand, deep learning approaches operate by having connected nodes that simulate human brains. The technique require huge amount of language resource and intensive computational resources as well from the technical point of view.

The main focus of the study is on the Amharic language, which is morphological rich, complex and a language without standard tools such as stemmers, PoS taggers, word nets, lexical dictionaries, or stop-word lists that are available for major languages such as English, French and Spanish. This presents a challenge to researchers working with the Amharic language.

Moreover, the complexity of the natural language applications may arise from character variations that generate same meaning from different orthographic representation of the given words. The character variation leads to a high number of independent variables which not only results in high computational complexity but also leads to learn in multiple meaning which affects the performance of the model.

The different character variation in the Amharic language produces the same meaning from alternative character representations of the given words, may contribute to the complexity of natural language applications. In addition to this, the Amharic language's character changes produce a high-dimensional feature space for machine learning, which not only results in a high level of computational complexity but is also are prone to what are called problems of 'overfitting'.

Because the Amharic language is morphologically rich, it is challenging to lemmatise textual data, although this is a desirable step for NLP. In addition, there are, as yet, no language resources that help to identify networks of synonyms and antonyms which in turn would allow developers to create generic, efficient and effective language tools. The Amharic language does not have sufficient and structured language corpora for the development of the natural language applications. To address this lack, we have attempted to collect and prepare text Amharic text corpora which can then be used for the development of different NLP applications. To check the usability of the developed text corpora, a document classification experiment is conducted using classical, ensemble machine learning as well as deep learning technique.

We recognize that this paper does not constitute applied corpus linguistics, that is, the application of corpus linguistics, but we offer it as an example of the problems involved in developing corpus resources for an under-resourced language, and of the means by which we have attempted to resolve this in the case of one language, Amharic. We intend that this paper can both document the procedures that we developed, and the means by which we, as NLP researchers, evaluated the resulting corpus, but also contribute to discussions of how large-scale corpus resources of under-resourced languages can be developed.

2. Motivation of this paper

As stated in the previous section, unlike the technologically favored languages, computational resource required for human language technology varies depending on the type of application and purpose, and these resources are not available for Ethiopian languages. These languages greatly suffer from the lack of language resources for NLP applications. The collection and preparation of corpora for Ethiopian languages specifically for the primary official working language of the country Amharic is, therefore, an important endeavor to facilitate the future research and development of NLP beside motivating the development of corpora for other Ethiopian languages.

In addition to this, the exponential growth in the number of news items that require a deeper understanding and text classification by topic has become an important issue for different applications using machine learning. We have, therefore, collected and prepared a total of 67,739 documents from eight news categories from Ethiopian online news sources. Currently, Amharic text news classification is done by journalists using a traditional approach which is time consuming, labor intensive and prone to error. This paper, therefore, proposes a new approach that involves collection and preparation of Amharic news corpora, and a series of experiments in news document classification that have the potential to alleviate these problems.

3. Related Works

Speech and text corpora have been developed as a basis for different natural language applications for a technological supported and resourced languages due to different reasons motivated by political and financial interest (Kurematsu, 1996; Suchomel, Pomikálek et al., 2012; Eggers, Malik, and Gracie, 2019). These technologically supported languages includes English, other European (such as French and Spanish) and Asian (Chinese and Japanese) languages.

By contrast, Ethiopia which contributes around 90 languages to the world, suffer from a lack of digital text and speech corpora for different natural language processing tasks (Eberhard and Fennig, 2020). All these languages are resource deficient and they are not technologically supported languages. These include but are not limited to Afaan Oromo, Amharic, Somali and Tigrigna languages which are the four most spoken languages in Ethiopia in the order they appeared (Sarah, 2019). To date, only a few attempts have been made towards developing limited speech and text corpora by different researchers (Adafre, 2004; Abate, Menzel, and Tafila, 2005; Argaw and Asker, 2005; Woldeyohannis, Besacier, and Meshesha, 2016; Abera and Hailemariam, 2018; Teshome, 2017; Abate et al., 2018a, 2020) .

Abate, Menzel, and Tafila (2005) created a phonetically rich and balanced read speech corpus that can be used for Amharic speech recognition using twenty hours of recordings from selected local news agencies. This speech corpus is still being used for developing speech recognition software. In addition, around eight hours of the Amharic read speech corpus has also been developed with the intention of translating speech from Amharic to English in tourism domain (Woldeyohannis, Besacier, and Meshesha, 2016). Thirdly, around twenty four hours of read speech corpus has also been developed for Tigrigna language using phonetically rich and balanced text from a news source (Abera and Hailemariam, 2018). Most recently, the development of speech corpora for Amharic, Afaan Oromo, Tigrigna and Wolaita languages contributed around twenty two hours of speech for each language (Abate et al., 2020).

The majority of the developed Amharic text corpora are intended for the purposes of machine translation (MT) compared to the other NLP applications. The first attempt to integrate Amharic into a unification based translation system was made by Adafre (2004) and this was followed by the development of Amharic to English and Tigrigna with the intent to create tools for translation between the respective languages (Woldeyohannis, Besacier, and Meshesha, 2016; Teshome, 2017). Very recently, another work have also been made towards developing a machine translation corpora between English and local Ethiopian languages as well as between Ethiopian languages which is made available in github[1] (Abate et al., 2018a,b). These parallel corpora collected are used for the development of automatic machine translation system in the absence of human intervention.

Beside the preparation of speech and machine translation data, in the attempt of Assefa and Goyal (2019), documents were collected by selecting four specific domains (Entertainment, Award, Telecommunication and Micro-credit industry) taking the data from eight web news sources which contributed towards the development of a document corpus from news. By contrast, other researchers have developed a text corpus ranging from four to sixteen domains

[1] Available at https://github.com/AAUThematic4LT/Parallel-Corpora-for-Ethiopian-Languages

(Eyassu and Gambäck, 2005; Asker et al., 2007; Teklu, 2012; Kelemework, 2013). However, the data in these cases were drawn from a limited range of news sources which seriously reduces the breadth of variation represented in the data.

In addition to this, a number of text and speech corpora have also been prepared by graduate students in their attempt to solve different tasks typically addressed in NLP including Question Answering (QA), Part-of-speech tagging (PoS), Document Classification, Named Entity Recognition (NER), Dialog System for different local languages. However, none of these data sets are available for use, for a number of reasons including the lack of a central repository, small size of data, limited number of domains, unbalanced data and the lack of domain experts involvement in the data preparation.

An additional desideratum is that, key considerations are that the corpus should be as generic as possible in supporting different NLP applications including Information Retrieval, Information Filtering, Sentiment Analysis, Recommendation System and Document Summarization and Classification (Kowsari et al., 2019). In addition to this, the data should support multiple domains but not be limited to Health, Social, Sport, Technology, Business, Law and Marketing. To this end, there should be at least some level of support for morphologically rich, complex and resource deficient languages regardless of other factors that make these languages overlooked by the state-of-the-art technology. Hence, in this paper, an attempt is made to report a generic approach to Amharic text preparation with reach and rich domain coverage, and a better size that can be used for different NLP applications.

4. Amharic Language

Amharic (**አማርኛ**) is spoken in Ethiopia since the late 12th century. Nowadays it is used as a means of communication by various sectors including the legal system, commerce, communications, the military and religion. According to Eberhard and Fennig (2020), the Amharic language is the second most spoken Semitic language in the world next to Arabic. Amharic is the primary working language of the Federal Government of Ethiopia and the regional languages including the Amhara, Diredawa and Southern Nations and Nationalities People Regions (SNNPR). Unlike other languages in Ethiopia, Amharic is also used as a tool for inter-regional communication.

According to Eberhard and Fennig (2020); Sarah (2019) and Thompson (2020), Amharic language is spoken by more than 27 million people with up to 22 million native speakers in Ethiopia having five dialectical variations spoken in different regions: Addis Ababa, Gojjam, Gonder, Wollo, and Menz. In other words, the majority of speakers are found in Ethiopia even though there are also speakers in different countries, particularly in Italy, Israel, Canada, the USA and Sweden.

Amharic belongs to Ethio-Semitic language groups derived from Ge'ez (**ግእዝ**) which is thought to be the historic center, classical and ecclesiastical language of Ethiopia (Yimam, 1986; Kogan, 2005). The language uses Ethiopic character derived from Geez with some added characters to fill the gap. Amharic has distinct features that make it different from other resourced languages such as English, European (Germany and Portuguese) and Asian (Mandrine and Japanese). These features include the alphabet, numbering system, gender sensitivity beside the complex nature of the phonetics, phonological and morphological properties (Samplius, 2020). Due to these characteristics, it is known to be morphologically rich and complex language. Sections 4.1 and 4.2 present the writing systems and morphological as well as syntactic features of Amharic languages, respectively.

4.1. Writing System

Amharic language uses a grapheme based writing system called fidel (**ፊደል**) Ge'ez script which is believed to date back to the 5th century BC. Fidel means "script", "alphabet", "letter", or "character". The writing system is also called Abugida (**አቡጊዳ**), taken from the first four symbols. The script is syllabic in which the character represents a combination of a consonant and a vowel, and the vowel is represented through modifications of the basic shape of the consonant.

The Amharic writing system has four distinct categories consisting of 276 distinct symbols: 231 core characters, 20 labiovelar symbols, 18 labialized consonants and 7 labiodental characters (Woldeyohannis and Meshesha, 2017). Table 1 below presents detail category of Amharic character set against their order.

	Character category			
	Core	**Labiodental**	**Labiovelar**	**Labialized**
Character	33	1	4	18
Order	7	7	5	1
Total	231	7	20	18

Table 1: Amharic character set distribution.

The first core characters possesses 33 primary characters and labiodental possess single character (ቭ/v/) each representing 7 orders, a consonant having one basic (ʔ/እ/) and six non-basic orders (əˈ/አ/, u/ኡ/, i/ኢ/, a/ኣ/, e/ኤ/, o/ኦ/) in the form to indicate the vowel which comes after the consonant to represent CV syllables. Unlike the core characters, a labiodental character only appears in modern loaned words borrowed from foreign languages like ቪዛ/visa/, ዩኒቨርሲቲ /University/ and ቫይረስ /virus/. Like the core characters, labiovelar category consists of 4 symbols (ቈ/k'^{w}/, ኈ/h^w/, ኰ/k^w/ and ጐ/g^w/) with 5 orders (əˈ/አ/, i/ኢ/, a/ኣ/, e/ኤ/, ʔ/እ/) that generates a total of 20 distinct graphemes. Likewise, labialized has 18 characters with single order (^{w}ä); this includes, ሏ/l^{w}ä/, ሟ/m^{w}ä/, ሯ/r^{w}ä/ and ሷ/s^{w}ä/ that are used in the Amharic writing system. In Amharic, there are 276 distinct symbols which are indispensable due to their distinct representation of the orthography.

In addition to the Ge'ez alphabets, Amharic graphemes also use Ge'ez number system in the publications of the official law documents, Negarit (government) magazine, Bible and other historic documents though the Arabic number is more frequently used and dominant in the modern literature (Fabri et al., 2014). However, in the publications of the official law documents, Negarit magazine, Bible and other historic documents use the Ge'ez number system more rather than the Arabic number. Table 2 depicts a sample of the Ge'ez and Arabic number systems.

Geez	፩	፪	፫	፬	፭	፮	፯	፰	፱	፲
Arabic	1	2	3	4	5	6	7	8	9	10
Geez	፳		፴		፵		፶		፷	
Arabic	20		30		40		50		60	
Geez	፸		፹		፺		፻		፲፻	
Arabic	70		80		90		100		1000	

Table 2: Geez and Arabic numerals used in Amharic writing (Foundation, 2020).

Moreover, the other distinguishing feature of Amharic is the punctuation mark used in the writing system of the language. The punctuation mark used in the Amharic orthography differs from that of the Latin-based writing system without any meaning difference in the writing systems. Like Amharic alphabets, Amharic punctuation marks are taken from the Ge'ez punctuation marks. This includes: Word separator (፡), Sentence end marker (።), Comma (፣) , Colon (፥), Semi-colon (፤), Preface-colon (፦), Paragraph separator (፨) and Question mark (፧). Among these punctuation, preface colon (፡), question mark (፧) and paragraph separator (፡) are no longer used in the modern Amharic writing system. In addition, word separator (፡) in the classical writing is replaced nowadays by white space because of the current practice of using computers in the Amharic document preparation without any change in meaning. Furthermore, the classical question mark (፧) has been replaced by the Latin question mark (?).

4.2. Morphological and Syntactic Features

Amharic makes use of the root and pattern system. The root (which is called a radical) is a set of consonants that bears the basic meaning of the lexical item and the pattern (Anbessa and Hudson, 2007; Leslau, 2000). The lexical item and the pattern is composed of a set of vowels inserted within the consonants of the root which resulted in derived words together with vowel. This derivation process makes Amharic language morphologically rich and complex.

Amharic languages follow Subject-Object-Verb (SOV) word-order(Leslau, 2000). For example, in the sentence **አበበ በሶ በላ።**/Abebe ate Beso./, the subject is **አበበ**/Abebe/, the object **በሶ**/Beso/, the verb is **በላ**/ate/and the sentence end marker **።** /./. The syntactic information of Amharic is also expressed at word level and the orthographic word may attach syntactic words like prepositions (**በ**/bä/, **ለ**/lä/ and **ከ**/kä/), connectors (**እና**/’əna/ and **ወይም**/wäyəmə/) and negation (**አል** /ʾälə/, **አይ**/ʾäyə/ and **አት**/ʾätə/) which create more than 100 word forms (Gasser, 2010, 2011). Beside these, verbs are inflected for tense, aspect, person, number, gender and mood while nominals are inflected for gender, number, definiteness and case (Griefenow-Mewis, 2001).

5. Data Collection and Preparation

The advancement of NLP has been transferred almost entirely to machine learning techniques than a direct software development, the whole process of NLP is powered by the respective data. The data used for NLP needs to be annotated in a way that reflect the real meanings of each statement so that it can be interpreted by the machine learning algorithms. The major and basic resource required for NLP applications are data in the form of large annotated, standard and representative text corpora. Such a corpus is either not available or inaccessible for Ethiopian languages in general and for Amharic in particular. Even the research attempted by different scholars had corpora of a small size, limited number of domain and imbalanced data for experiment purpose which in turn negatively or adversely affect the result obtained from the experiment. Thus, the collection and preparation of labeled, standard and representative text corpora for Amharic language is an important endeavor for the future development of NLP application in Amharic.

Consequently, the following sections briefly discusses the method of data collection in section 5.1 followed by pre-processing for data cleaning, Amharic character normalization and punctuation correction in section 5.2; and corpus size distributions in section 5.3.

5.1. Data Collection

The research team has applied different techniques to collect a general purpose text corpus for natural language processing from more than 25 registered news and religious domain of different sources which provides the freely accessible news in different language and categories. The sources are selected based on the language of the news, ability to provide the news in electronic format and a minimum of three years in the area. This local news includes: Fana Broadcasting Corporate (FBC[2]), Addis Admas News (AAN[3]), Ethiopian Press Enterprise (EPE[4]), Ethiopian Orthodox Tewahido Church (EOTC[5]), Ethiopian Orthodox Tewahido Church Mahibere Kidusan (EOTCMK[6]), Ethiopian Reporter Amharic (ERA[7]), Ethiopian News Agency (ENA[8]) and Walta Information Center (WIC[9]) which provides news in the form of text in machine readable format. From these news sources a total of 67,739 documents were collected to the end of October, 2020. Table 3 presents the detail of document collected against the source and category class.

During document preparation, it is observed that the minimum number of sentences are limited to 2 per document with up to a maximum number 871 sentences. The reason is that, as we collect corpus from online sources a significant number of documents has a title as a header information without the news content in text forms. The header information is followed by the audio or video which is out of the coverage of this study. The text data collected from news sources include predefined eight main categories. These main categories

[2] https://www.fanabc.com/
[3] https://www.addisadmassnews.com/
[4] https://www.press.et/
[5] https://www.ethiopianorthodox.org/
[6] https://eotcmk.org/a/
[7] https://www.ethiopianreporter.com/
[8] https://www.ena.et/
[9] http://www.waltainfo.com/index.php?locale=am

	AAN	FBC	WIC	Religion	ENA	ERN	EPE	Total
Social	6,008	-	2,956	-	11,459	4,027	440	24,890
Politics	332	-	4,622	-	6,697	3,821	400	15,872
Business	693	720	2,398	-	7,234	3,970	532	15,547
Religious	-	-	-	3,949	-	-	-	3,949
Sport	-	327	887	-	-	1,277	836	3,327
Health	353	107	329	-	1,064	-	179	2,032
Sci-Tech	-	128	421	-	651	247	81	1,528
Law	-	-	-	-	-	540	54	594
Total	7,386	1,282	11,613	3,949	27,105	13,882	2,522	67,739

Table 3: Amharic documents collected by the source and category class.

are: business, health, religion, social, law, politics, science as well as technology and sport. The news category selected based on the content that each online news provides as a link in consultation with journalist expert. For instance, the sport category includes the subcategories of football and athletics in one source of news content and local and worldwide sports in another. Similar subcategories of the social category include culture, recreation, tradition, and others.

As depicted in table 3, most sources organize news items as per the defined different categories except law and religion. However, some news item do not categorize their information under the areas of law and religion separately. They are categorized as the miscellaneous area because they do not produce a lot of material for these domains. These caused the bulk of the categories for law and religion to be empty.

To extract a web news item from the different websites, a web crawler is used for each article after identifying the structure of web documents (html) including the page navigation. Some news item uses same web document structure while others do not which requires a different algorithm to crawl the web news. Accordingly, in this paper web structural analysis and extraction of the content is done.

5.2. Preprocessing

Data pre-processing is the first step towards natural language processing for a better result and understanding. The majority of available text data is highly unstructured, which is not easy for machines to understand. The problems of unstructured data include typographic errors, concatenated words, bad grammar, abbreviations, idiomatic expression, usage of slang and presence of unwanted white space and emojis.

As part of the preprocessing, we performed Amharic characters normalization, sentence segmentation and data cleaning for the removal of unnecessary links, emoji, symbols and foreign words as well as the removal of the extra white space incorporated in the web documents. To perform the pre-processing tasks, Python scripts were used including the libraries such as natural language processing toolkit (NLTK) and Regular Expressions (RE). The following section presents details about the Amharic character normalization and Punctuation correction.

Normalization

Amharic text contains some characters that have similar roles and are effectively redundant. These characters do not bear different meaning when one is replaced with the other characters in the modern Amharic writing system though they possess semantic differences in the traditional writings. These redundant characters are (**ሀ** /hə/) which can also be written as (**ሀ** /hə/, **ሃ** /ha/, **ሐ** /hä/, **ሓ** /ha/, **ኀ** /hä/ and **ኃ** /ha/), **ሰ** /sə/ can be written as **ሰ** /sə/ or **ሠ** /sə/, **አ** /ˀə/ can be written as, **አ** /ˀə/ or **ዐ** /ˀə/ and **ፀ** /s'ə/ can be written as **ፀ** /s'ə/ and **ጸ** /s'ə/ along with their orders.

In the current Amharic writing system, these characters are used interchangeably. As a result, one Amharic word sentence is summed up to provide a large number of word variants without meaning difference as discussed in (Abate et al., 2018a). For example, consider a sample English sentence "today i will take my son to the zoo": its equivalent translate to Amharic sentence, "**ልጄን ዛሬ ወደ እንስሳት መንከባከቢያ እወስደዋለሁ**" /ləǧenə zare wädä 'ənəsəsatə mänəkäbakäbiya 'əwäsədäwalähu/ which generate a total of 128 different sentences as a result of እ/'ə/, ስ/sə/, ሳ /sa/ and ሁ/hu/. To avoid words or sentences with the same meaning of different orthographic representation from being taken as different, we have replaced a set of characters with the most frequently used character in Amharic document to greatly minimize ambiguity in natural language processing.

Punctuation Correction

Data collected from the web contain formats with single and double quotes. In addition, the classical writing system has a number of punctuation marks that are either not used in the modern writing system or replaced with other characters. These are: word separator (፡ **ሁለት ነጥብ**) /hulätə nät'əbə/, sentence end marker (። **አራት ነጥብ**) /ᵓäratə nät'əbə/, comma (፣), colon (፥ **ነጠላ ሰረዝ**) /nät'äla säräzə/, semi-colon (፤ **ድርብ ሰረዝ**) /dərəbə säräzə/, preface-colon (፦ **ሁለት ነጥብ ከሠረዝ**) /hulätə nät'əbə kəsäräzə/), paragraph separator (፨ **አራት ነጥብ**) /ᵓäratə nät'əbə/ and question mark (፧ **ትእምርተ ጥያቄ**) /tə'əmərətä t'əyaqe/). Among these punctuation marks, preface colon (፦), question mark (፧) and paragraph separator (፨) are no longer used in the modern Amharic writing system. The question mark (፧) is replaced by the Latin question mark (?) and paragraph separator (፨) is also replaced by the sentence end marker (።). In addition, the word separator (፡) used to separate words in classical writing is replaced in the current practice with white space without any change in meaning. Accordingly, we replaced all punctuation marks from their usage in the classical writing with those used in modern writing in Amharic documents.

5.3. Corpus Size and Data Distribution

The data distribution demonstrates how well the data fits into a certain model. The more data there is, the higher chance the machine learning algorithm has in comprehending it and creating accurate predictions for the unknown data. As a result, this part provides information about data distribution in terms of document, sentence, and word (tokens[10] and types[11]) beside the source of news and domain.

The corpora have been preprocessed, normalized and analyzed to see the distribution of the content against the different categories and online news sources. The extracted corpus have a total of 67,739 documents consisting of 1,795,320 sentences with a total of 26,515,769 word tokens and 883,339 word types. The collected corpora has a minimum of three tokens and types per sentence with a minimum of three sentences per documents. In addition to this, the maximum of 789 tokens and 572 types has been registered in the collected documents with an average of 15 tokens and types per sentence beside an average of 27 sentences per documents. Figure 1 presents the detail distribution of the Amharic document collected per news source.

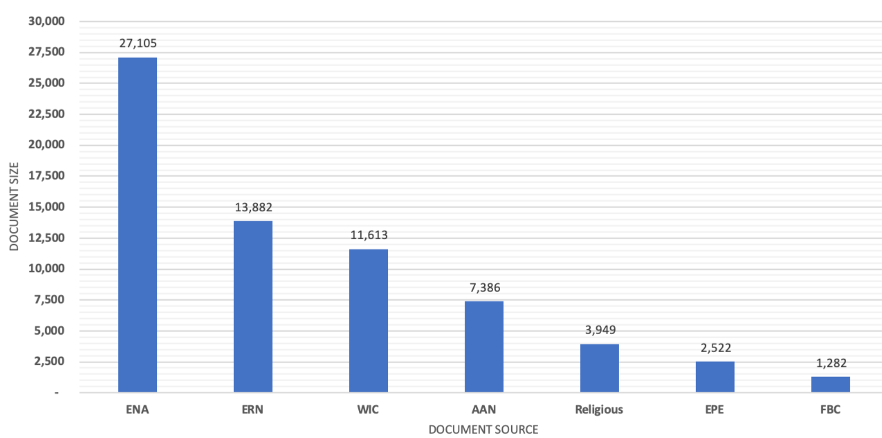

Figure 1: Amharic document collected per news source

As depicted in Figure 1, among all the data sources, Ethiopian News Agency (ENA) contributes around 40% of the total Amharic documents. This is because it is one of the state owned and main news agency. Ethiopian Reporter

[10]Token refers to the total number of words in the collected document regardless of repetition.

[11]Type refers to the total number of distinct words in the collected document without any repetition.

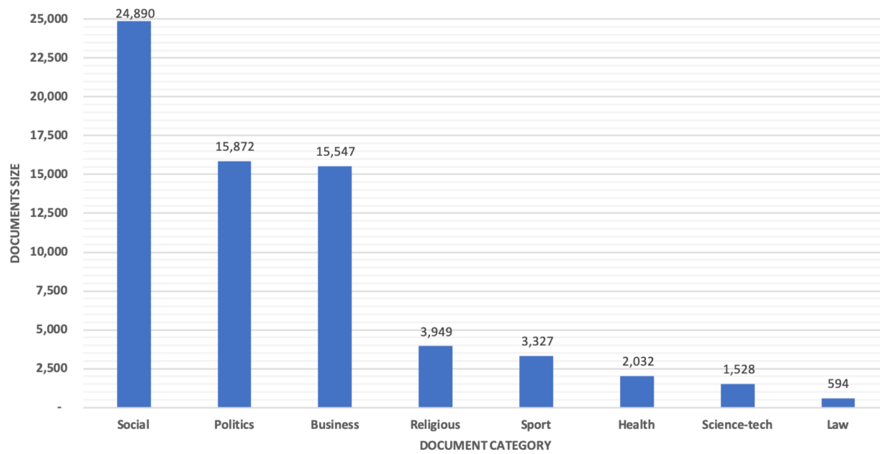

Figure 2: Amharic document collected for each category.

News (ERN) and Walta Information Center (WIC) come second and third with 20.49% and 17.14% contribution to the total Amharic document collection. By contrast, Fana Broadcast Corporate (FBC) contributes around 1.89% of the total data which is the smallest number of documents. The main reason behind collecting small size documents on FBC is that, the news agency makes the video data more available than a text data. Similarly, Figure 2 depicts the distribution of Amharic documents collected for each category.

It is observed that most of the Amharic documents collected from online sources are in social category which accounts for 36.74% of the total collection. This is followed by 23.43% documents in the politics and 22.95% documents in the business category. One of the reason for a large number of document in social category is that it contains the number of documents that intersect with different categories. These sub categories include art, entertainment and culture which can not be distinguished from the social category. Hence, this category is used to organize documents containing social and related issues.

Conversely, from the very nature of Amharic news, law category contributes less than 1% documents which is available only in two news sources from the total eight different online news sources while science and technology contributes around 2.26% documents though it is collected from five different news sources. Amharic language has more complex morphology than other resourced and technologically supported languages. To show the morphological richness and complexity of the Amharic languages, the distribution of tokens, types, sentences and number of documents are presented for each category. Table 4 presents the detail distribution of token, type and sentence per each category.

	Token	Type	Sentence	Document
Business	5,799,933	257,603	360,631	15,547
Health	687,256	79,322	48,892	2,032
Law	1,103,562	132,414	68,093	595
Politics	5,647,034	265,300	359,634	15,872
Religious	2,214,120	194,746	152,236	3,949
Science-Technology	373,093	50,671	26,814	1,528
Social	9,478,170	534,702	693,836	24,890
Sport	1,212,601	104,332	85,184	3,327

Table 4: Token, type, sentence and document size of Amharic document for each document category.

As depicted in Table 4, in terms of the distribution of tokens and sentences, the social, business and politics category data dominates while law is the first in terms of data richness because of the large text coverage in type, token and sentence despite the smaller size of documents. Even though health as well as science and technology contain a higher number of documents than law, they are not rich in terms of tokens, types and sentences. In addition, the data distribution of tokens, types and sentences are presented taking the minimum, maximum and average distribution of the collected data for each category to show the richness. Accordingly, table 5 presents the minimum, maximum and average number of tokens, types and sentences per class.

	Token			Type			Sentence		
	Min	Max	Mean	Min	Max	Mean	Min	Max	Mean
Business	30	3,883	373	27	1,965	255	3	493	24
Health	18	2,747	339	17	1,386	237	3	338	24
Law	157	6,107	1,858	139	2,836	1,130	10	565	115
Politics	16	8,053	356	9	3,766	242	3	871	23
Religious	15	6,833	561	13	3,369	385	3	371	39
Science-Tech	22	1,419	244	20	993	179	4	455	18
Social	16	4,380	381	15	2,601	269	3	520	28
Sport	27	2,476	365	25	1,429	254	3	645	26

Table 5: The minimum, maximum and average number of word token, word type and sentence per category.

As presented in table 5, the collected document collections contain from 15 to 8,053 tokens, 9 to 3,766 types and 3 to 871 sentences. On average, the documents prepared contains per sentence 15 tokens and 14 types with an average of 27 sentences per document.

6. Experimental Results

Since the major challenges facing Amharic NLP research has been the lack of standardized, domain rich, balanced and large size text corpora, an attempt has been made here to collect, prepare and pre-process data collected from online sources. In this paper the text corpora is checked and validated for its usefulness using Amharic text classification using different machine learning algorithm. Furthermore, the experiment assists the system in automatically categorizing the given Amharic text news into one of the pre-defined categories. Accordingly, section 6.1 below discusses the details of experimental setup for Amharic document classification while section 6.2 presents the experimental results and finding of the study.

6.1. Experimental Setup

The Amharic document classification conducted using a total of 67,739 documents from eight different categories. The corpus is divided into two datasets: training and testing dataset. The training dataset contains 80% of the documents used for constructing classification model, while the remaining 20% are reserved as test set for evaluating the classification model. The data is partitioned uses the default random state setting as per the recommendation of different researchers in Amharic document classification and the power law distribution (Bar-Yam, 2016).

To validate the usability of the prepared corpora, a document classification experiment are conducted by selecting six machine learning algorithms. Accordingly, Naive Bayes (NB) and Support Vector machine (SVM) from classical machine learning (Zheng, 2019), and Gradient Boosting (GB) and Random Forest (RF) from ensemble learning are used (Onan, Korukoğlu, and Bulut, 2016), while Deep Neural Network (DNN) and Convolutional Neural Network (CNN) from deep-learning technique are employed (Kim, 2014; Conneau et al., 2016). The selection of these machine learning techniques is based on their best performance on the document classification as per different research reports (Keyvanpour and Imani, 2013; Goudjil et al., 2018; Liu et al., 2017).

The final part of the text classification is to evaluate the performance of the model. There are a number of methods available for evaluating supervised machine learning being accuracy is the simplest method (Huang and Ling, 2005). Accuracy is calculated by finding the ratio of number of correct prediction to the total number of test dataset.

6.2. Experimental Results and Discussion

Table 6 presents the detail of experimental results to validate the usability of the corpus for Amharic document classification.

Among all the baseline experiments of Amharic document classification, as presented in Table 6, the performance accuracy registered for deep learning shows higher than any other classical (SVM and NB) and ensemble based (GB

	Classification		Accuracy
	Correct	Incorrect	
Support Vector Machine (SVM)	10,601	2,947	78.25%
Naive Bayes (NB)	10,493	3,055	77.46%
Gradient Boosting (GB)	10,130	3,418	74.78%
Random Forest (RF)	10,300	3,248	76.03%
Convolutional Neural Network (CNN)	11,208	2,340	82.73%
Deep Neural Network (DNN)	11,451	2,097	84.53%

Table 6: Experimental result for the Amharic document classification.

and RF) machine learning techniques. Among the deep learning, DNN which is special type of Recurrent Neural Network (RNN) has presented the highest performance than that of CNN. DNN has improved the accuracy of document classification by 1.8% than that of CNN with a relative error reduction of 10.42% (82.73% to 84.53%). One of the reason for the high performance by DNN over CNN is that, the convolutions and pooling operations works by selecting the best performers. This makes the CNN to ignore the local ordering of words which makes it harder to classify taking the sequence of words which is not in DNN (Kim, 2014).

Compared to classical machine learning, a Gradient Boosting classifier using ensemble techniques shows the lowest performance than any other experiment conducted in ensemble and deep learning techniques as depicted in Table 6. Gradient Boosting improved by relative error rate of 4.96% (74.78% to 76.03%) in using Random Forest and at most 38.66% (74.78% to 84.53%) while using DNN.

Similarly, the experimental result of Naive Bayes (NB) registered a lower performance from the classical machine learning than the deep learning techniques. The amount of relative error reduced by at least 3.50% (77.46% to 78.25%) and at most 31.36% (77.46% to 84.53%) by using Support Vector Machine (SVM) and Deep Neural Network (DNN) respectively. The reason why a better performance was registered in SVM than NB is that, NB treat the text data as independent features while SVM attempt to look at the interaction to certain extent. In addition to this, NB performs better in snippets than that of the full-length document (Wang and Manning, 2012). Figure 3 depicts the baseline performance measures of the classical, ensemble and deep learning technique for the Amharic document classification.

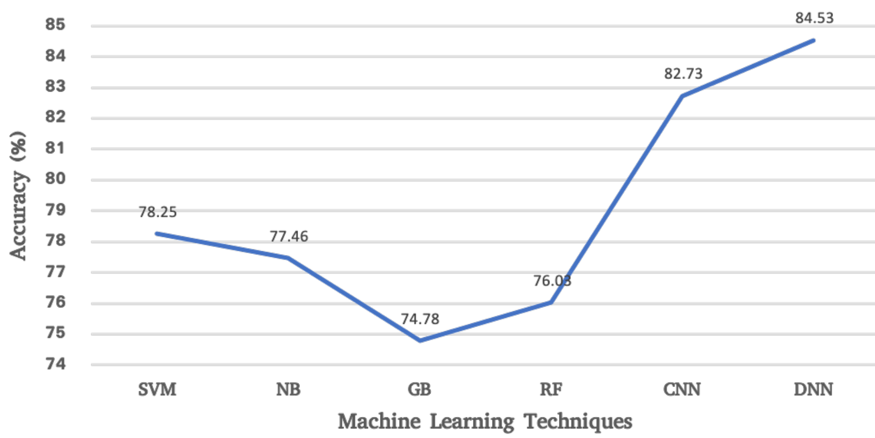

Figure 3: Performance measure of the machine learning techniques in Amharic document classification.

The baseline experiment of the Amharic document classification shows that the corpus collected from news and religious source can be used with at least 74% accuracy for the document classification. Compared to the experiment conducted in Eyassu and Gambäck (2005); Asker et al. (2007); Kelemework (2013); Tegegnie, Tarekegn, and Alemu (2017), the result registered in this experiment is promising given the data size, data variety, data complexity, concept usage variety and content semantics of the corpus organized for different NLP applications.

In general, the results obtained in this experiment are promising. As the main aim of this paper is to produce large size corpora, rich in content with a variety of data, the result shows the usability of the Amharic document corpus which can serve as a test bed for designing natural language applications. It is a well-known fact that as the size and variety of data increases, it brings lexical, syntactic, semantic and discourse ambiguity for natural language application, and hence calls for different researchers to design and develop NLP applications.

7. Concluding Remarks

This paper presents an attempt to collect and prepare usable Amharic text corpora for one of the Ethiopian languages, Amharic. Corpora have been collected from eight different sources including the religious website using the state-of-the-art Python libraries. The collected corpora are further processed for performing normalization, punctuation correction and data cleaning in the course of preparing the corpus for different NLP tasks. The primary NLP applications that the corpora can be used for are: automatic speech recognition, machine translation, plagiarism detection, topic modeling, text summarization and information retrieval. In addition to this, the collected corpora can contribute towards the development of NLP resources including stop-word list, word sense disambiguation, relation extraction, stemming and lemmatization, language model, part of speech tagger, morphological segmentation, analysis and synthesis.

To check the usability of the collected, preprocessed and normalized data, the Amharic text document classification experiments have been conducted using machine learning techniques and promising results were achieved. The experimental results further show that the corpus collected can be used for designing and developing a number of NLP applications and resources. As the source for the corpus data is news, there is a need to update the corpus continually and make it available for researchers preferably through a third party who may show an interest. In addition, the complexity of the data collected both in the size and variety calls for more research to solve the challenge of lexical, syntactic and semantic ambiguities so as to develop more efficient and effective Amharic NLP applications and resources.

This paper has demonstrated an approach to the development of a corpus for a resource-deficient language, Amharic, for use in NLP applications. While the paper is not directly reflective of the application of corpus linguistics, it contributes to the discussion of how resources and tools can be created for corpus analysis of languages that are lacking in such resources.

References

Abate, Solomon Teferra, Michael Melese, Martha Yifiru Tachbelie, Million Meshesha, Solomon Atinafu, Wondwossen Mulugeta, Yaregal Assabie, Hafte Abera, Binyam Ephrem Seyoum, Tewodros Abebe, et al. 2018a. Parallel corpora for bi-lingual english-ethiopian languages statistical machine translation. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3102–3111.

Abate, Solomon Teferra, Michael Melese, Martha Yifiru Tachbelie, Million Meshesha, Solomon Atinafu, Wondwossen Mulugeta, Yaregal Assabie, Hafte Abera, Binyam Ephrem, Tewodros Abebe, Wondimagegnhue Tsegaye, Amanuel Lemma, Tsegaye Andargie, and Seifedin Shifaw. 2018b. Parallel corpora for bi-directional statistical machine translation for seven Ethiopian language pairs. In *Proceedings of the First Workshop on Linguistic Resources for Natural Language Processing*, pages 83–90, Association for Computational Linguistics, Santa Fe, New Mexico, USA.

Abate, Solomon Teferra, Wolfgang Menzel, and Bairu Tafila. 2005. An Amharic speech corpus for large vocabulary continuous speech recognition. In *Ninth European Conference on Speech Communication and Technology*.

Abate, Solomon Teferra, Martha Yifiru Tachbelie, Michael Melese, Hafte Abera, Tewodros Abebe, Wondwossen Mulugeta, Yaregal Assabie, Million Meshesha, Solomon Afnafu, and Binyam Ephrem Seyoum. 2020. Large vocabulary read speech corpora for four ethiopian languages: Amharic, Tigrigna, Oromo and Wolaytta. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 4167–4171.

Abera, Hafte and Sebsibe Hailemariam. 2018. Design of a tigrinya language speech corpus for speech recognition. In *Proceedings of the First Workshop on Linguistic Resources for Natural Language Processing*, pages 78–82.

Adafre, Sisay Fissaha. 2004. *Adding Amharic to a Unification-based Machine Translation System: An Experiment*. Peter Lang, Bern, Switzerland.

Anbessa, Teferra and Grover Hudson. 2007. Essentials of Amharic.

Argaw, Atelach Alemu and Lars Asker. 2005. Web mining for an Amharic-english bilingual corpus. In *WEBIST*, pages 239–246.

Asker, Lars, Atelach Alemu Argaw, Björn Gambäck, and Magnus Sahlgren. 2007. Applying machine learning to Amharic text classification. In *Proceedings of the 5th World Congress of African Linguistics*.

Assefa, Misrak and Vishal Goyal. 2019. Amharic text news classification. *Journal of Emerging Technologies and Innovative Research*.

Bar-Yam, Yaneer. 2016. Concepts: power law. *New England Complex Systems Institute. Retrieved*, 21.

Bender, Emily M and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Bergstra, James, Olivier Breuleux, Frédéric Bastien, Pascal Lamblin, Razvan Pascanu, Guillaume Desjardins, Joseph Turian, David Warde-Farley, and Yoshua Bengio. 2010. Theano: a cpu and gpu math expression compiler. In *Proceedings of the Python for scientific computing conference (SciPy)*, volume 4, pages 1–7, Austin, TX.

Chawla, Nitesh V, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. 2002. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357.

Conneau, Alexis, Holger Schwenk, Loïc Barrault, and Yann Lecun. 2016. Very deep convolutional networks for text classification. *arXiv preprint arXiv:1606.01781*.

Eberhard, Gary F. Simons, David M. and Charles D. Fennig. 2020. *Ethnologue: Languages of the World*. SIL International, Dallas, Texas.

Eggers, William D, Neha Malik, and Matt Gracie. 2019. Using ai to unleash the power of unstructured government data. *Deloitte Insights*.

Eyassu, Samuel and Björn Gambäck. 2005. Classifying Amharic news text using self-organizing maps. *43rd Annual Meeting of the Association for Computational Linguistics; Workshop on Computational Approaches to Semitic Languages*.

Fabri, Ray, Michael Gasser, Nizar Habash, George Kiraz, and Shuly Wintner. 2014. Linguistic introduction: The orthography, morphology and syntax of semitic languages. In *Natural Language Processing of Semitic Languages*. Springer, pages 3–41.

Foundation, Ge'ez Frontier. 2020. Digital resources for the realm of ge'ez.

Gasser, Michael. 2010. A dependency grammar for Amharic. In *Proceedings of the Workshop on Language Resources and Human Language Technologies for Semitic Languages, Valletta, Malta*.

Gasser, Michael. 2011. Hornmorpho: a system for morphological processing of Amharic, Oromo, and Tigrinya. In *Conference on Human Language Technology for Development, Alexandria, Egypt*.

Goudjil, Mohamed, Mouloud Koudil, Mouldi Bedda, and Noureddine Ghoggali. 2018. A novel active learning method using svm for text classification. *International Journal of Automation and Computing*, 15(3):290–298.

Griefenow-Mewis, Catherine. 2001. *A grammatical sketch of written Oromo*, volume 16. Rüdiger Köppe.

Huang, Jin and Charles X Ling. 2005. Using auc and accuracy in evaluating learning algorithms. *IEEE Transactions on knowledge and Data Engineering*, 17(3):299–310.

Jurafsky, Daniel and James H Martin. 2008. Speech and language processing: An introduction to speech recognition, computational linguistics and natural language processing. *Upper Saddle River, NJ: Prentice Hall*.

Kelemework, Worku. 2013. Automatic Amharic text news classification: A neural networks approach. *Ethiopian Journal of Science and Technology*, 6(2):127–137.

Keyvanpour, Mohammad Reza and Maryam Bahojb Imani. 2013. Semi-supervised text categorization: Exploiting unlabeled data using ensemble learning algorithms. *Intelligent Data Analysis*, 17(3):367–385.

Kim, Yoon. 2014. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*.

Kogan, Leonid. 2005. Common origin of ethiopian semitic: the lexical dimension. *Scrinium*, 1(1):367–396.

Kowsari, Kamran, Kiana Jafari Meimandi, Mojtaba Heidarysafa, Sanjana Mendu, Laura E. Barnes, and Donald E. Brown. 2019. Text classification algorithms: A survey. *Information*, 10(4).

Kurematsu, Akira. 1996. *Automatic Speech Translation*, volume 28. CRC Press.

Leiner, Barry M, Vinton G Cerf, David D Clark, Robert E Kahn, Leonard Kleinrock, Daniel C Lynch, Jon Postel, Larry G Roberts, and Stephen Wolff. 2009. A brief history of the internet. *ACM SIGCOMM Computer Communication Review*, 39(5):22–31.

Leslau, Wolf. 2000. *Introductory grammar of Amharic*, volume 21. Otto Harrassowitz Verlag.

Li, Xigen. 2013. *Internet newspapers: The making of a mainstream medium.* Routledge.

Liu, Jingzhou, Wei-Cheng Chang, Yuexin Wu, and Yiming Yang. 2017. Deep learning for extreme multi-label text classification. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 115–124.

Meyer, Ronny. 2006. Amharic as lingua franca in ethiopia. *Lissan: Journal of African Languages and Linguistics*, 20(1/2):117–132.

Onan, Aytuğ, Serdar Korukoğlu, and Hasan Bulut. 2016. Ensemble of keyword extraction methods and classifiers in text classification. *Expert Systems with Applications*, 57:232–247.

Park, Namkee, Jae Eun Chung, and Seungyoon Lee. 2012. Explaining the use of text-based communication media: An examination of three theories of media use. *Cyberpsychology, Behavior, and Social Networking*, 15(7):357–363.

Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830.

Sammut, Claude and Geoffrey I Webb. 2011. *Encyclopedia of machine learning.* Springer Science & Business Media.

Samplius. 2020. The main characteristics of amharic language.

Sarah, Adem. 2019. The most spoken languages in ethiopia 2019. [Online; accessed 25-June-2021].

Suchomel, Vít, Jan Pomikálek, et al. 2012. Efficient web crawling for large text corpora. In *Proceedings of the seventh Web as Corpus Workshop (WAC7)*, pages 39–43.

Tegegnie, Alemu Kumilachew, Adane Nega Tarekegn, and Tamir Anteneh Alemu. 2017. A comparative study of flat and hierarchical classification for Amharic news text using svm. *International Journal of Information Engineering & Electronic Business*, 9(3).

Teklu, Surafel. 2012. *Automatic categorization of Amharic news text: a machine learning approach*. LAP Lambert Academic Publishing.

Templeton, Graham. 2020. Why Natural Language Processing (NLP) Needs Natural Language Data. https://www.globalme.net/blog/collect-natural-language-data-for-nlp/. [Online; accessed 25-June-2021].

Teshome, Mulu Gebreegziabher. 2017. *English-Amharic Statistical Machine Translation*. Ph.D. thesis, PhD Dissertation, IT Doctoral Program, Addis Ababa University, Addis Ababa, Ethiopia. Unpublished thesis.

Thompson, Irene. 2020. About world language. Accessed: 2021-06-26.

Wang, Sida I and Christopher D Manning. 2012. Baselines and bigrams: Simple, good sentiment and topic classification. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 90–94.

Woldeyohannis, Michael Melese, Laurent Besacier, and Million Meshesha. 2016. Amharic speech recognition for speech translation. *JEP-TALN-RECITAL*, page 114.

Woldeyohannis, Michael Melese and Million Meshesha. 2017. Experimenting statistical machine translation for ethiopic semitic languages: The case of Amharic-Tigrigna. In *International Conference on Information and Communication Technology for Develoment for Africa*, pages 140–149, Springer.

Yimam, Baye. 1986. Amharic grammar. *Addis Ababa: EMPDA*.

Zheng, Yuhan. 2019. An exploration on text classification with classical machine learning algorithm. In *2019 International Conference on Machine Learning, Big Data and Business Intelligence (MLBDBI)*, pages 81–85, IEEE.