Evaluating the Effectiveness of the Standard Insights Extraction Pipeline for Bantu Languages

Mathibele Nchabeleng and Joan Byamugisha(✉)

IBM Research Africa, Johannesburg, South Africa
{mathibele.nchabeleng,joan.byamugisha}@ibm.com

Abstract. Extracting insights from data obtained from the web in order to identify people's views and opinions on various topics is a growing practice. The standard insights extraction pipeline is typically an unsupervised machine learning task composed of processes that preprocess the text, visualize it, cluster and identify the topics and sentiment in each cluster, and then graph the network. Given the increasing amount of data being generated on the internet in Africa today, and the multilingual state of African countries, we evaluated how well the standard pipeline works when applied to text wholly or partially written in indigenous African languages, specifically Bantu languages. We carried out an exploratory investigation using Twitter data and compared the outputs from each step of the pipeline for an English dataset and a mixed Bantu language dataset. We found that for Bantu languages, due to their complex grammatical structure, extra preprocessing steps such as part-of-speech tagging and morphological analysis are required during data cleaning, threshold values should be adjusted during topic modeling, and semantic analysis should be performed before completing text preprocessing.

Keywords: Insights extraction · Bantu languages · Twitter data

1 Introduction

The growing penetration of mobile telephony and internet services in Africa has led to an increased presence of African user-generated content, especially on social media platforms (such as Facebook, Twitter, and WhatsApp). According to Internet World Stats [8], by the end of 2018, over 460 million out of the continent's 1.3 billion people used the internet, and there were over 200 million Facebook subscribers at the end of 2017. This represents a 35.2% internet penetration rate and a 15.5% Facebook penetration rate [8]. The user-generated content has been leveraged to obtain insights about elections [23], design marketing strategies [1], and monitor the aftermath of epidemics [19]. However, only the content that is written in languages with high-quality linguistic resources

J. M. Jose et al. (Eds.): ECIR 2020, LNCS 12035, pp. 159–172, 2020.
https://doi.org/10.1007/978-3-030-45439-5_11

such as English, French, Portuguese, and Arabic are used for such analyses and content generated in indigenous African languages is largely excluded.

It has been found that even though the amount of content generated in indigenous African languages is significantly lower than non-indigenous language content, it nonetheless contains valuable insights, especially relevant to the local context [12]. Hence, it is extremely important that we develop resources and tools that can be used to parse out useful information from free-text written in any language. In this paper, we investigated whether the standard insights extraction pipeline is sufficient when applied to a single language family indigenous to Africa, Bantu languages, using the following questions: (1) how well does the standard insights extraction pipeline apply to Bantu languages; and (2) if found to be inadequate, why, and how can the pipeline be modified so as to be applicable to Bantu languages?

Two datasets of 20,000 tweets each were included in the study: one was comprised solely of English text and the other a mixed batch of six Bantu languages and English text. Both datasets were analysed using a seven-step pipeline: (1) text preprocessing and normalization, (2) dimensionality reduction, (3) visualization, (4) clustering, (5) topic modeling, (6) sentiment analysis, and (7) network graphing; and the differences in outcomes were measured.

We found that: (1) there is a need to differentiate between conjunctively and disjunctively written languages; (2) sentiment analysis should be performed before verb stemming during text preprocessing, before any present negation morpheme is removed; (3) during text preprocessing and normalization, stemming verbs and adjectives is crucial to avoiding very high levels of sparsity in the representation matrix; (4) stemming nouns must be avoided so as to prevent the loss of important semantic information; and (5) during topic modeling, some threshold values must be adjusted to account for agglutination. This evaluation has, to the best of our knowledge, never been done for Bantu languages.

The rest of the paper is arranged as follows: in Sect. 2, a brief background on Bantu languages and their grammatical structure is presented; Sect. 3 presents related work on extracting insights using the standard pipeline; and the methods, investigation, and results of the evaluation are presented in Sect. 4. The implications of our findings are discussed in Sect. 5, and we conclude in Sect. 6.

2 Brief Background on Bantu Languages

Bantu languages are indigenous to Africa, geographically extending from the south, below Nigeria, to most of central, east, and southern Africa, they are found in 27 of the continent's 54 countries, and range in number from 300 to 680 [21]. Bantu languages have an agglutinating morphology, where words consist of several morphemes, and each affix agglutinated with the root word carries meaning such as tense and aspect [21]. The writing system of Bantu languages is either conjunctive or disjunctive [25]. In the former case, several orthographic words, 'I love them', are written as a single word, for example, *Mbakunda* in Runyankore (a language indigenous to Uganda). The latter case writes different orthographic words as separate words. For example, the same translation for 'I love them' is *Kea ba rata* in Sepedi (a language indigenous to South Africa).

The hallmark of Bantu nominal morphology is the noun class (NC), where all nouns are assigned to a class; and there are over 20 NCs, although some have fallen into disuse in most languages [17,21]. A simple noun comprises a prefix and a stem [11]; for example, *omuntu*, 'person' in Runyankore, can be analyzed as the prefix *o-mu-* and stem *-ntu*. However, not all Bantu languages have the initial vowel on the prefix [11,17]; for example, 'person' in Sepedi is *motho*, with prefix *mo-* and stem *-tho*. Noun classes are also at the heart of an extensive system of concordial agreement that governs grammatical agreement in verbs, adjectives, possessives, subject, object, etc. [11,25]; this is a pivotal constituent of the whole Bantu sentence structure [25].

The morphological and phonological structure of Bantu verbs is very regular in most languages [20,24], with a typical verbal form consisting of: one or more bound morphemes, a verb-root, and one or more extensions [24]. The morphemes preceding the verb-root specify the person, noun class, tense, aspect, time, negation, etc., while the extensions specify valency-changing categories,—the arguments controlled by a verb—which can be as many as eleven [24]. Additionally, Bantu languages typically have a large number of tenses, with up to four observed past tenses, and up to three observed discrete future tenses [20].

This complex grammatical structure is partly what has led to Bantu languages being largely computationally under-resourced, despite still primarily being a first language throughout the continent. In the next section, we present related work on extracting insights from collections of data.

3 Standard Insights Extraction Pipeline

Social media data mining has become a common tool used to extract opinions from a large population in order to monitor, understand, and predict people's reactions to an event, and to measure the diffusion of ideas within the social network [15]. In this section, related work on extracting insights from collections of documents is presented. The scope here is limited to social media data, specifically textual data, and more so, Twitter data, because the vast amount of content generated and shared through social media contains rich knowledge and covers a wide spectrum of social dynamics [33]. In their socio-semantic analysis of Twitter data, Lipizzi et al. [15] stated that the following processes are necessary to extract complete and valuable insights from data: (1) preprocessing the text, (2) identifying and classifying opinions in the network, (3) analyzing the sentiment of individual or groups of text, (4) visualization of the large amounts of data; and (5) extracting conversational maps from social streams. We subdivided these processes into seven steps, including dimensionality reduction and clustering, shown in Fig. 1.

Text preprocessing is necessary because of the strong heterogeneity and noisiness characteristic of social media texts [26]. It involves dealing with incorrect spelling, contractions and abbreviations, stop words, inflectional variants, user

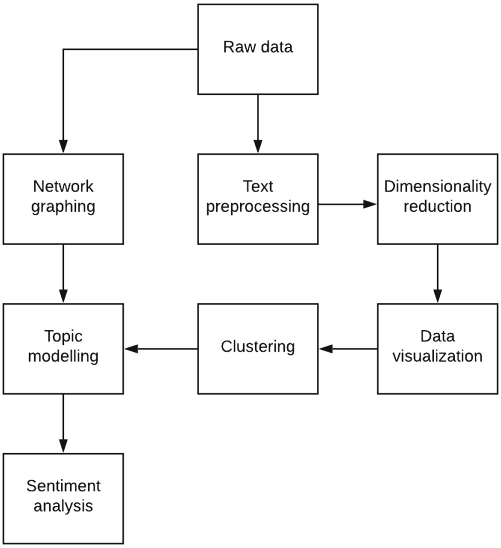

Fig. 1. The seven steps for extracting insights from social media data

tags, hyperlinks, numbers, and email addresses [26,32]. However, the steps taken during preprocessing depend on the quality, quantity, and style of the underlying text [32].

Data visualization requires that the large amount of data be compacted in an expressive fashion [15]. Because text documents are high dimensional objects, effectively visualizing such data requires it to be projected to a lower-dimensional space [18]. Thus, dimensionality reduction, which involves transforming high-dimensional data into a meaningful representation of reduced dimensionality, is an essential part of text mining [6,31]. However, for the lower dimensional representation to be meaningful, it must be a good approximation of the original document set given in its full space [6]. The commonly used techniques for dimensionality reduction are: Principal Component Analysis (PCA), which both minimizes information loss and increases interpretability [6,10]; Singular Value Decomposition (SVD), a stable and robust technique [9]; and T-Distributed Stochastic Neighbor Embedding (T-SNE), which is particularly well suited for the visualization of high dimensional datasets [30].

Document clustering aims to efficiently organize, navigate, summarize, and retrieve documents [3]. It can either be done using partitioning algorithms, where the number of clusters is specified before clustering takes place (for example, K-Means [2,4], Locally Adaptive Clustering [3]), and Non-negative Matrix Factorization [13,14,29]; or hierarchical algorithms, which start by either considering each document as a cluster (agglomerative clustering) or all documents as belonging to a single cluster (divisive clustering) [2,35]. In the former case, documents are continually assigned to the nearest cluster until no further improvement is achieved, while the latter either decreases or increases the number of clusters until a stopping condition is met [2,35].

Topic modeling is the application of probabilistic models to uncover the underlying semantic structure of a collection of documents, where each topic is defined as a distribution over a set of words [2,34]. There are several topic modeling algorithms, but the most commonly applied are Latent Dirichlet Allocation (LDA) [2,4,26,33,34] and Non-negative Matrix Factorization (NMF) [7,13,14,22,29].

Sentiment analysis involves the computational study of people's opinions, appraisals, attitudes, and emotions about events, entities, individuals, and topics [16,27]. Features found to be important during sentiment analysis include: terms and their frequency, adjectives, negation, and opinion words and phrases [16]. Sentiment analysis can be formulated either as a supervised learning problem that can be solved using well-known classification algorithms such as Naive Bayes or Support Vector Machines [16,27], or as an unsupervised problem where opinion words and phrases are used as the dominating indicators of sentiment [16].

Network graphing is used to provide structure to the information exchanged in a social network, and has mostly been used to identify influential users on a topic for marketing or advertising services [4]. Here, each user in a social network is considered as a node in a graph, and the relationships between users (follow, retweet, like, etc.) as directed edges between nodes in the graph [4].

4 Evaluation of Suitability of Pipeline to Bantu Languages

The above processes have been found to be sufficient to extract insights from text in other languages beyond English, such as French [26], Chinese [34], and Arabic [2]. However, to the best of our knowledge, no work has been done completely to apply the described pipeline to Bantu languages. Here, the methodology and results of evaluating the suitability of the standard pipeline for use with Bantu languages are presented.

4.1 Materials and Methods

We used two datasets in this evaluation, each comprising 20,000 tweets; the first, an English dataset composed of customer reviews[1]; the second, composed of tweets in English, mixed code, and six Bantu languages, was archived directly from live South African and Ugandan tweets covering the period February 2019 to May 2019. The live tweets were archived based on the trending hashtags during the period of data collection. The six Bantu languages targeted were IsiZulu, Luganda, Runyankore, Sepedi, Sesotho, and Setswana. These languages were

[1] The English dataset is available from https://www.kaggle.com/thoughtvector/customer-support-on-twitter.

selected because they cover both conjunctive and disjunctive writing styles, and they are understood by the authors. However, due to the use of the mixed code writing style, we found tweets that contained terms in other Bantu languages beyond the six considered.

Our investigation was limited to Twitter data due to the inherent difficulty of performing opinion mining on it, resulting from the informal writing style used and limited tweet length. We hypothesize that the findings based on Twitter data are generalizable to other social media platforms. We further limited the size of each dataset to 20,000 tweets, as the results based on a limited dataset are also generalizable to a larger dataset. Both datasets were run through the seven processes in the standard pipeline, and analyzed for any significant differences. For text preprocessing, we used the same techniques as described in [26,32]. However, no stemming/lemmatizing was performed on either dataset because, to the best of our knowledge, two of the Bantu languages (Luganda and Runyankore) do not have tools for this[2]. We used multiple approaches for dimensionality reduction (PCA, and a combination of SVD followed by T-SNE), clustering (K-Means and NMF) and topic modeling (LDA and NMF), in order to consider the approach which gives the better result. Gephi[3] was used to graph the network.

At each step in the pipeline, the results between the two datasets were compared, with emphasis placed on any observed differences, significant or otherwise. Where a process in the pipeline was found to be insufficient to process the Bantu language dataset, we then investigated if and how the complex grammatical structure of these languages causes the observed limitations. We further investigated what needs to be done in order to adapt that process to fulfill the same task for Bantu languages. On the other hand, where a process in the pipeline was found to adequately apply to the Bantu language dataset, we noted this finding and proceeded to the next step.

4.2 Results

At the end of the evaluation, the processes of text preprocessing, topic modeling, and sentiment analysis were found to require some modification in order to sufficiently extract meaningful insights from textual data in Bantu languages. The processes of dimensionality reduction, data visualization, and clustering, though being language independent, were also found to be affected by the term-document matrix, which is itself language dependent. Only network graphing was found to be completely language independent. The following subsections provide details on the limitations found during text preprocessing, topic modeling, and sentiment analysis, and explain the findings based on the grammatical structure of Bantu languages.

[2] This fact is true for the majority of Bantu languages, and is yet another example of their under-resourced state.

[3] Gephi is available from https://gephi.org/users/download/.

Text Preprocessing. During text preprocessing, after converting the data to lower case, it underwent the removal of HTML tags, URLs, numbers, email addresses, Twitter handles, and hashtags; then the expansion of contractions (such as *can't* and *we're*) and abbreviations (such as *lol*, *dm*, and *tbh*); and finally, the elimination of non-alphanumeric characters and stop words. The text in both datasets was not stemmed or lemmatized due to the lack of such resources for some of the Bantu languages considered in this investigation.

For the English-only dataset, the preprocessing performed was found to be sufficient. However, we found that several additional processes are necessary to fully preprocess the mixed Bantu language dataset. These processes are: distinguishing conjunctively versus disjunctively written languages, part-of-speech tagging, and stemming/lemmatizing only verbs and adjectives.

Distinguishing Between Conjunctively and Disjunctively Written Languages. The mixed Bantu language dataset comprised three conjunctively written languages (isiZulu, Luganda, and Runyankore) and three disjunctively written languages (Sepedi, Sesotho, and Setswana). As explained in Sect. 2, Bantu languages are written either conjunctively or disjunctively, and therefore, there is a need to differentiate between them in order to perform the appropriate preprocessing. Taljard and Bosch [25] identified that a word-class tagger is sufficient for disjunctively written languages, while a morphological analyzer is required for the conjunctively written languages. This is because the disjunctive system of writing requires bound morphemes to be written as orthographically distinct units (*Kea ba rata* 'I love them' in Sepedi), thus making morphological information explicit in the orthography [25]. On the other hand, the conjunctive writing style requires a morphological analyzer to make the different morphemes in the orthography explicit [25], for example from *Mbakunda* to *m-ba-kunda* 'I love them' in Runyankore. The authors concluded that the differences in writing systems necessitate the use of different architectures specifically for part-of-speech tagging. The need for part-of-speech tagging was identified as crucial during text preprocessing and therefore, the type of writing style first needs to be identified before this can be performed.

Part-Of-Speech Tagging. Though neither stemming nor lemmatization were performed on both datasets during preprocessing, we nonetheless recognize the need to stem/lemmatize the verbs and adjectives because of their numerous grammatical forms. Nouns, on the other hand, should not be stemmed as this would result in the loss of their core semantics. As explained in Sect. 2, a noun is composed of a prefix and a stem. However, the stem of a noun is not unique, but rather gets its full semantics from the prefix. Table 1 shows examples of tweets from the dataset where stemming the noun will result in a meaningless stem[4] (the nouns of interest are in bold font, with the prefix underlined).

Table 1. The noun stems *-pedi*, *-tswana*, and *-ntu* are not unique and are meaningless on their own

Language	Tweet	Stem
Sepedi	*o* ***mopedi*** *empa o palela kego ngwala* ***sepedi***	*-pedi*
Setswana	*south african* ***setswana*** *eseng* ***setswana*** *sa* ***botswana***	*-tswana*
isiZulu	***umuntu*** _ng_***umuntu*** _ng_***abantu***	*-ntu*

The examples shown in Table 1 highlight the problem that can result if nouns are stemmed during text preprocessing. For Sepedi, *sepedi* (a language) would be indistinguishable from *mopedi* (a member of the Bapedi tribe); for Setswana, *setswana* (a language) would have the same stem as *Botswana* (a country); for isiZulu, *umuntu* (person) would be reduced to the same stem as *abantu* (people). Additionally, for isiZulu, a conjunctively written language, the example also shows the need for morphological analysis, to separate the copulative *ng* from the noun.

With the semantics of the noun removed through stemming, the resultant stems *-pedi*, *-tswana*, and *-ntu* are meaningless without a prefix. This in turn would affect topic modeling downstream. Part-of-speech tagging is therefore required to differentiate between nouns that should not be stemmed and other parts-of-speech that should.

Stemming Verbs and Adjectives. A typical Bantu language verbal form consists of one or more bound morphemes, a verb-root, and one or more extensions [24]. The bound morphemes include the subject and object, which are determined by the noun class, as is the full adjectival form [11,25]. Therefore, for a language like Runyankore with 20 noun classes, there are 400 different ways of conjugating a single verb stem for subject and object. Additionally, the number of extensions can be as many as nine, as shown in Table 2, where a single verb stem *reeb-* in Runyankore and *bon-* in Sepedi for 'see' is extended.

In addition to the increasing number of verb forms owing to the extensions shown in Table 2 and the noun class system, Bantu languages typically have a very large number of tenses [20]. For example, Runyankore has 14 tenses [28] and these too are part of the verb form. This complex grammatical structure results in a single verb root having thousands of possible verb forms. Therefore, verb stemming/lemmatizing is a crucial step during preprocessing, which, if not performed results in very high levels of sparsity in the resultant matrix.

[4] The translations to the text in Table 1 are:
Sepedi: O mopedi empa o palela kego ngwala sepedi, 'Your native tongue is Sepedi but you can't even write the language'
Setswana: South African Setswana eseng Setswana sa Botswana, 'South African Setswana not the Setswana from Botswana'
isiZulu: Umuntu ngumuntu ngabantu, 'A person is a person through/because of (other) people'.

Table 2. Different verb extensions for the verb stems *reeb-* in Runyankore and *bon-* in Sepedi (the dashes between the letters represent the separation between the verb root and the extensions)

Runyankore	Sepedi
Reeb-a (See)	*Bon-a* (See)
Reeb-er-a (See for)	*Bon-a-ng* (See, both or all of you)
Reeb-erer-a (Look after)	*Bon-a-ne* (See each other, must)
Reeb-w-a (Seen by)	*Bon-a-le* (Be visible, must)
Reeb-an-a (Look at each other)	*Bon-a-na* (See each other)
Reeb-ek-a (Materialize)	*Bon-a-la* (Be visible)
Reeb-uur-a (Observe)	*Bon-a-la-ng* (Who/which shows)
Reeb-agur-a (Stare)	*Bon-a-gala* (Become visible)
Reeb-a-reeb-a (Look around)	*Bon-a-gala-go* (Who/which become(s) visible)
Reeb-es-a (See with)	*Bon-a-la-go* (Who/which are visible)

Adjectives also require stemming because the full form of an adjective depends on the noun class of the noun being described. Therefore, the number of forms that a single adjective can take depend on the number of noun classes in that language. Runyankore, for example, has 20 different forms for each adjectival stem because it has 20 noun classes. Table 3 shows some examples of the forms that the adjective 'beautiful' in Runyankore *-rungi* and Sepedi *-botse* (the adjective prefix is underlined).

Table 3. The different adjectival forms for the stems *-rungi* in Runyankore and *-botse* in Sepedi

English	Runyankore	Sepedi
Beautiful woman	*Omukazi murungi*	*Mosadi yo mobotse*
Beautiful children	*Abaana barungi*	*Bana ba ba botse*
Beautiful guava	*Eipeera rirungi*	*Kwaba ye botse*
Beautiful eyes	*Amaisho marungi*	*Mahlo a mabotse*
Beautiful building	*Ekizimbe kirungi*	*Moago o mobotse*
Beautiful leg	*Okuguru kurungi*	*Leoto le lebotse*

Topic Modeling. Topic modeling was performed using LDA and NMF. In the case of the mixed Bantu language dataset, we used all tweets and tokens during the modeling. The average tweet length in the corpus by Dela Rosa et al. [5] was 15.22, so we considered a very low threshold of at least five tokens per tweet. From these datasets, it was found that 21.26% of tweets in the mixed Bantu language dataset were below this threshold, compared to 16.22% in the English dataset. While this difference is not significant, we emphasize the agglutinative structure of Bantu languages presented in Sect. 2, where a word consists of several morphemes, and each affix agglutinated with the root word carries meaning such as subject, object, tense, aspect, negation, etc. [21]. As a result, for the conjunctive writing style, an entire sentence can be represented as a single word. Consider the following Runyankore example from [28]: *Titukakimureeterahoganu*, meaning 'We have never ever brought it to him', and comprises the morphemes *ti-tu-ka-ki-mu-reet-er–a-ho-ga-nu*. For this reason, all tweets, despite their length, were included in the topic modeling of the mixed Bantu language dataset.

We also included all tokens in the mixed Bantu language dataset during topic modeling. Although, this is contrary to the recommended minimum token count of three, it was done because, as explained in Sect. 4.2, a single verb stem can be inflected into thousands of verb forms, and it should therefore be expected that, without performing verb stemming, such tokens will be extremely rare in the dataset. From measuring the number of tokens below the recommended threshold count of three in both datasets, we found that 72.04% of tokens in the mixed Bantu language dataset were below this threshold, compared to 0.00% in the English dataset. This is a significant result, again pointing to the importance of verb stemming during preprocessing. Conversely, the English dataset, which was not stemmed either, does not show such an adverse need for it.

Table 4. Negation in Runyankore and Sepedi (the negation morphemes are underlined)

Conjunctive (Runyankore)	Disjunctive (Sepedi)
*oru runyankore nanye tinaru**kyeng**a*	*ka sepedi ga re **berekis**e c q z le x*
*tiha**ri**ho border erikwatanitsa uganda na rwanda*	*ga re **buw**e sesotho mo limpopo*
*konkashi eki otarikuki**reeb**a noha*	*ka sepedi bare tshwene ga e **ipon**e makopo*

Sentiment Analysis. There are currently no publicly available sentiment analysis implementations for any of the Bantu languages used during this investigation. However, we assessed the currently available tools to evaluate whether sentiment analysis could be done following the standard pipeline. In Sect. 3, four features were identified as important for sentiment analysis; three of these (terms and their frequency, adjectives, and opinion words and phrases) are also applicable to a Bantu language dataset. However, if verb stemming is performed during text preprocessing (as we recommend in Sect. 4.2), negation will present a differentiating factor for Bantu languages. This is because, for conjunctively written languages, the negation morpheme(s) is agglutinated to the verb stem, while for disjunctively written languages, the negation morpheme is not necessarily only used in the context of negation. Consider the excerpts from the dataset shown in Table 4[5] writing styles (the negation morphemes are underlined and the verb roots are in bold font).

In the standard pipeline, sentiment analysis is performed after text preprocessing, visualization, clustering, and topic modeling, in order to assess the sentiment associated within a specific cluster or topic. However, for Bantu languages, once verb stemming is performed during text preprocessing, then the verbs in Table 4 are reduced to their roots (shown in bold font); thus losing the negation morphemes *ti*, *ta*, *ga*, and *e*.

Further complexities during sentiment analysis arise from:(1) multiple rules regarding negation, and (2) the negation morpheme being applicable to other parts of speech other than negation. For the former case, consider the example of Runyankore, where *ti* is the primary negative and *ta* the secondary negative; Sepedi, in addition to the negation morpheme *ga*, encodes negation in the change of the final vowel from *a* to *e*. Losing such morphemes would in turn skew the results on sentiment analysis further down the pipeline.

5 Discussion

From the findings presented in Sect. 4, we have shown that Bantu languages require a different architecture from the ‘standard’. We therefore propose an alternative architecture shown in Fig. 2.

The following are the areas where differences arise (note that the other processes maintain their original placement in Fig. 1):

(1) During text preprocessing, identifying the writing style of a language is done first, to determine whether to perform part-of-speech tagging for Disjunctively written languages or morphological analysis for conjunctively written languages.
(2) Next, part-of-speech tagging and morphological analysis are performed to prevent nouns from being stemmed, thus avoiding the loss of their semantics encoded in the noun prefix, and ensure that verbs and adjectives are stemmed in order to avoid noise in the data and high levels of sparsity in the resultant matrix.
(3) Sentiment analysis is performed during text preprocessing, before any negation morphemes are lost during verb stemming. Further, it is performed after part-of-speech tagging and morphological analysis, in order to avoid the ambiguity of the negation morpheme identified for some disjunctively

[5] The translations to the text in Table 4 are:
Oru Runyankore nanye tinarukyenga, ‘I have also not understood this Runyankore’
Tihariho border erikwatanitsa Uganda na Rwanda, ‘There is no border that joins Uganda and Rwanda’
Konkashi eki otarikukireeba noha, ‘But honestly, who does not see this’
Ka sepedi ga re berekise c q z le x, ‘In sepedi, we don’t use the letters c, q, z, and x’
Ga re buwe sesotho mo limpopo, ‘We don’t speak Sesotho in Limpopo’
Ka sepedi bare tshwene ga e ipone makopo, ‘In Sepedi, they say, “A monkey does not see its own forehead”.’.

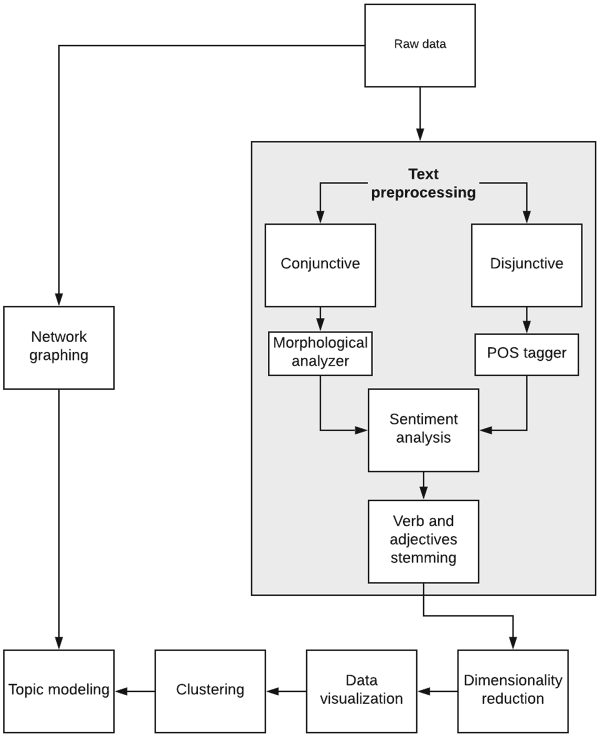

Fig. 2. The revised insights extraction pipeline for Bantu languages

written languages, while also making the negation morpheme explicit for conjunctively written languages.

(4) Finally, during topic modeling, without stemming verbs and adjectives, the threshold counts should not be applied because a significant amount of the dataset will be excluded.

6 Conclusion

In this paper, the standard insights extraction pipeline was evaluated for how well it applies to a grammatically complex and under-resourced family of languages, Bantu languages. Seven processes were identified as belonging to the standard pipeline (text preprocessing, dimensionality reduction, visualization, clustering, topic modeling, sentiment analysis, and network graphing) and tested for their effectiveness on two datasets of 20,000 tweets each, one composed of English and the other a mixture of English and six Bantu languages. Results showed that: conjunctively written languages should be distinguished from disjunctively written languages, because they require different preprocessing steps; verbs and adjectives, but not nouns, should be stemmed; threshold counts should be revised during topic modeling; and sentiment analysis should be done before verb stemming, in order to prevent the loss of the negation morpheme. Future work will include implementing these recommendations and assessing their effectiveness.

References

1. Afolabi, A.: Social Media Marketing: The Case of Africa. Master's thesis, Carleton University, Ontario, Canada (2016)
2. Alhawarat, M., Hegazi, M.: Revisiting k-means and topic modeling: a comparison study to cluster arabic documents. IEEE Access **6**, 42740–42749 (2018)
3. AlSumait, L., Domeniconi, C.: Text clustering with local semantic kernel. In: Berry, M.W., Castellanos, M. (eds.) Survey of Text Mining, Clustering, Classification and Retrieval, 2nd edn, pp. 87–105. Springer, London (2007). https://doi.org/10.1007/978-1-84800-046-9_5
4. Cha, Y., Cho, J.: Social network analysis using topic models. In: 35th Annual SIGIR Conference (SIGIR 2012), pp. 565–574. ACM, Portland (2012)
5. Dela Rosa, K., Shah, R., Lin, B., Gershman, A., Frederking, R.: Topical clustering of tweets. In: Proceedings of the ACM SIGIR: SWSM, vol. 63 (2011)
6. Howland, P., Park, H.: Cluster preserving dimension reduction methods for document classification. In: Berry, M.W., Castellanos, M. (eds.) Survey of Text Mining, Clustering, Classification and Retrieval, 2nd edn, pp. 3–23. Springer, London (2007). https://doi.org/10.1007/978-1-84800-046-9_1
7. Hoyer, P.O.: Non-negative matrix factorization with sparsity constraints. J. Mach. Learn. Res. **5**, 1457–1469 (2004)
8. Internet World Stats: Africa Internet user stats in 2019 population by country (2019). https://internetworldstats.com/africa.htm. Accessed 11 Apr 2019
9. Jackson, J.E.: A Users Guide to Principal Components Analysis. Wiley, New York (1997)
10. Jolliffe, I.T., Cadima, J.: Principal component analysis: a review and recent developments. Philos. Trans. R. Soc. A. Math. Phys. Eng. Sci. **374**(2065), 20150202 (2016)
11. Katamba, F.: Bantu nominal morphology. In: The Bantu Languages: Routledge Language Family Series, vol. 4, chap. 7, pp. 103–120. Taylor and Francis/Routledge, London (2003)
12. Kende, M., Quast, B.: Promoting content in Africa (2016). https://www.internetsociety.org/wp-content/uploads/2017/08/Promoting20Content20In20Africa.pdf. Accessed 11 Apr 2019
13. Kim, J., Park, H.: Sparse nonnegative matrix factorization for clustering. Technical report, Georgia Institute of Technology (2008)
14. Kuang, D., Choo, J., Park, H.: Nonnegative matrix factorization for interactive topic modeling and document clustering. In: Celebi, M.E. (ed.) Partitional Clustering Algorithms, pp. 215–243. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-09259-1_7
15. Lipizzi, C., Iandoli, L., Ramirez-Marquez, J.E.: Extracting and evaluating conversational patterns in social media: a socio-semantic analysis of customers' reactions to the launch of new products using Twitter streams. Int. J. Inf. Manag. **35**, 490–503 (2015)
16. Liu, B., Zhang, L.: A survey of opinion mining and sentiment analysis. In: Aggarwal, C., Zhai, C. (eds.) Mining text data, pp. 415–463. Springer, Boston (2012). https://doi.org/10.1007/978-1-4614-3223-4_13

17. Maho, J.: A comparative study of bantu noun classes. Ph.D. thesis, Goteborg University, Goteborg, Sweden (1999)
18. Mao, Y., Balasubramanian, K., Lebanon, G.: Dimensionality reduction for text using domain knowledge. In: 23rd International Conference on Computational Linguistics (COLING 2010), pp. 801–809. Association for Computational Linguistics (2010)
19. Morin, C., Most, I., Mercier, A., Dozon, J.P., Atlani-Duault, L.: Information circulation in times of Ebola: Twitter and the sexual transmission of Ebolaby survivors. PLoS Currents **10** (2018)
20. Nurse, D.: Aspect and tense in bantu languages. In: The Bantu Languages: Routledge Language Family Series, vol. 4, chap. 6, pp. 90–102. Taylor and Francis/Routledge, London (2003)
21. Nurse, D., Philippson, G.: Introduction. In: The Bantu Languages: Routledge Language Family Series, vol. 4, chap. 1, pp. 1–9. Taylor and Francis/Routledge, London (2003)
22. Peharz, R., Stark, M., Purnkopf, F.: Sparse non-negative matrix factorization using l0-constraints. In: IEEE International Workshop on Machine Learning for Signal Processing. IEEE, Kittila (2010)
23. Portland Africa: How Africa tweets 2018 (2018). https://portland-communications.com/publications/how-africa-tweets-2018/. Accessed 08 Feb 2019
24. Schadeberg, C.T.: Derivation. In: The Bantu Languages: Routledge Language Family Series, vol. 4, chap. 5, pp. 71–89. Taylor and Francis/Routledge, London (2003)
25. Taljard, E., Bosch, S.: A comparison of approaches to word class tagging: conjunctively versus disjunctively written Bantu languages. Nord. J. Afr. Stud. **15**, 428–442 (2006)
26. Tapi-Nzali, M.D., Bringay, S., Lavergne, C., Mollevi, C., Opitz, T.: What patients can tell us: topic analysis for social media on breast cancer. J. Med. Internet Res. (JMIR) **5**(3), e23 (2017)
27. Thakkar, H., Patel, D.: Approaches for sentiment analysis on Twitter: a state-of-art study. Computer Science, Social and Information Networks. arXiv:1512.01043 (2015)
28. Turamyomwe, J.: Tense and Aspect in Runyankore-Rukiga: Linguistic Resources and Analysis. Master's thesis, Norwegian University of Science and Technology, Norway (2011)
29. Túrkmen, A.C.: A review of non-negative matrix factorization methods for clustering. Stat.ML. arXiv:1507.03194v2 (2015)
30. van der Maaten, L., Hinton, G.E.: High-dimensional data using t-SNE. J. Mach. Learn. Res. **9**, 2579–2605 (2008)
31. van der Maaten, L., Postma, E., van der Herik, H.: Dimensionality reduction: a comparative review. Tiburg Centre for Creative Computing, Tilburg University, Technical report (2009)
32. Wesslen, R.: Computer assisted text analysis for social science: topic models and beyond. Computation, and Language, Computer Science (2018)
33. Wu, Y., Cao, N., Gotz, D., Tan, Y.P., Kim, D.A.: A survey on visual analytics of social media data. IEEE Trans. Multimedia **99**(1) (2016)
34. Wu, Y., Ding, Y., Wang, X., Xu, J.: A comparative study of topic models for topic clustering of Chinese web news. In: 3rd IEEE International Conference on Computer Science and Information Technology (ICCSIT), vol. 5 (2010)
35. Xu, R., Wunsch, D.: Survey of clustering algorithms. IEEE Trans. Neural Netw. **16**(3), 645–678 (2005)